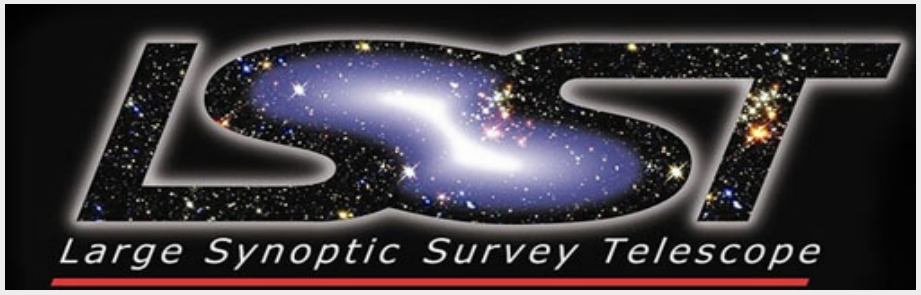


Mining Citizen Science Data: Machine Learning Challenges

Kirk Borne
George Mason University
kborne@gmu.edu , http://classweb.gmu.edu/kborne/

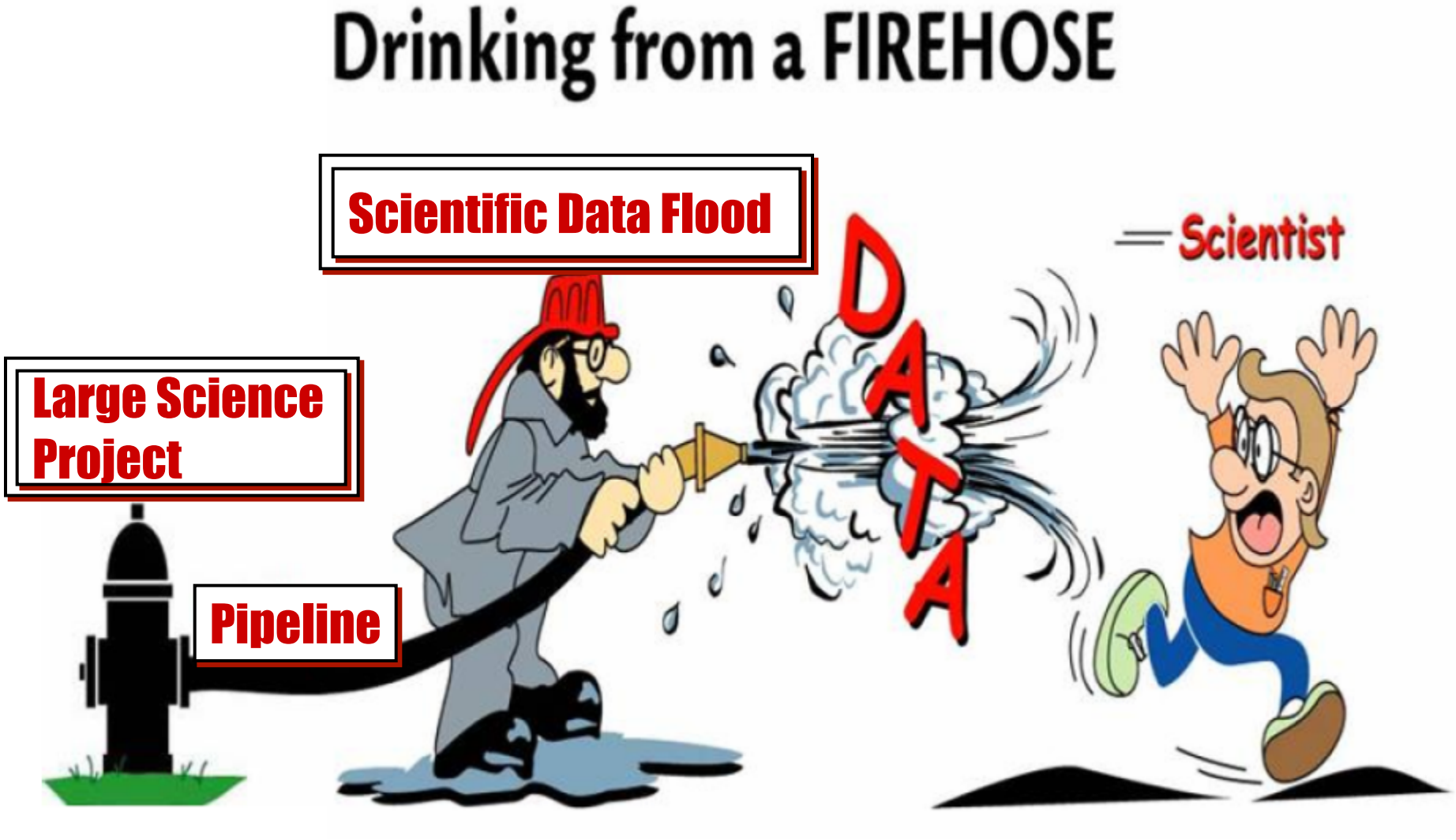
Outline

- The Emergence of Informatics Research in the Sciences
- Human Computation
- Citizen Science: Introduction
- Citizen Science: Examples
- The Zooniverse Project
- The Zooniverse Machine Learning Challenges

Outline

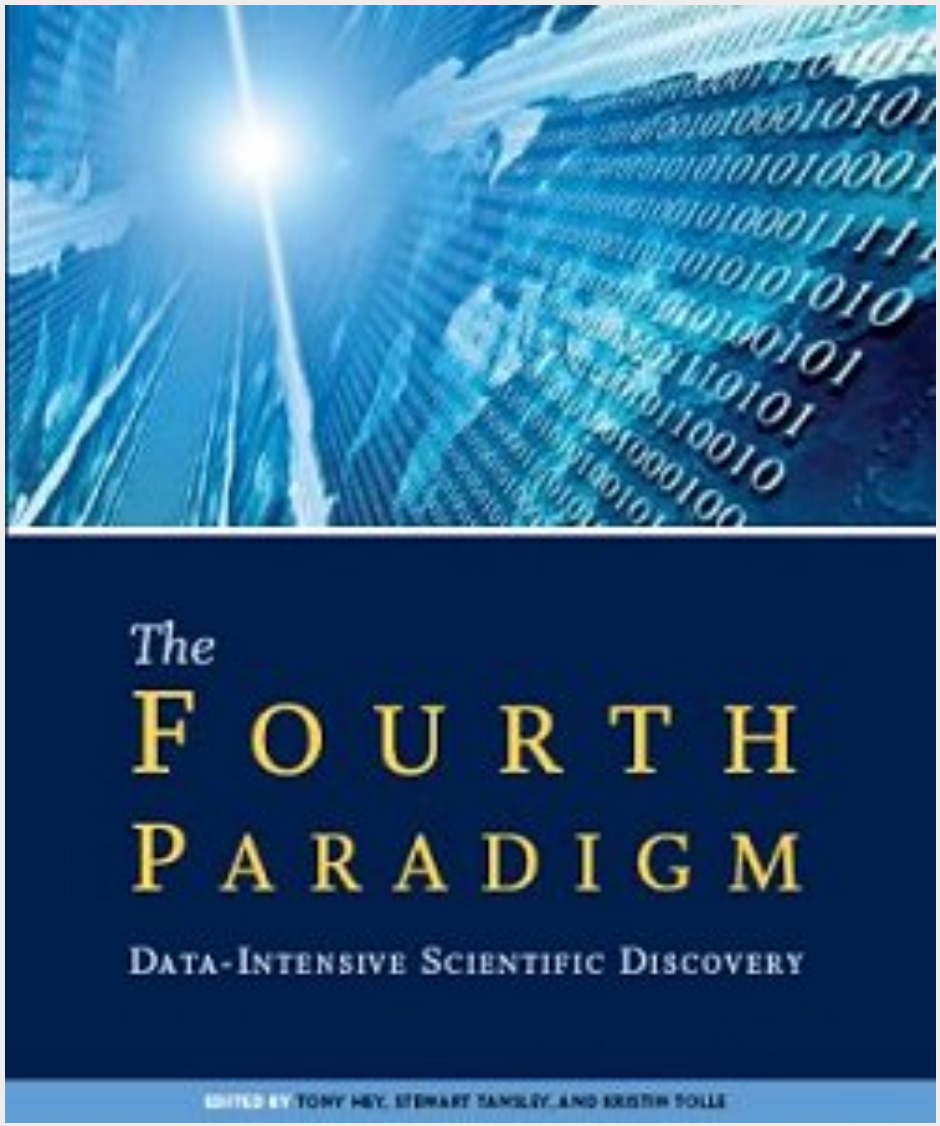
- **The Emergence of Informatics Research in the Sciences**
- Human Computation
- Citizen Science: Introduction
- Citizen Science: Examples
- The Zooniverse Project
- The Zooniverse Machine Learning Challenges

The Scientific Data Flood



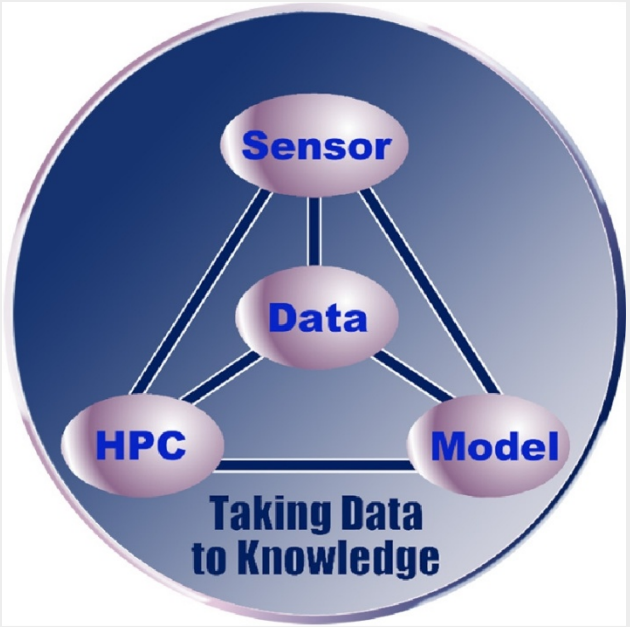
The Fourth Paradigm: Data-Intensive Scientific Discovery

http://research.microsoft.com/en-us/collaboration/fourthparadigm/



The 4 Scientific Paradigms:

1. Experiment (sensors)
2. Theory (modeling)
3. Simulation (HPC)
4. **Data Exploration (KDD)**



General Themes in Informatics Research

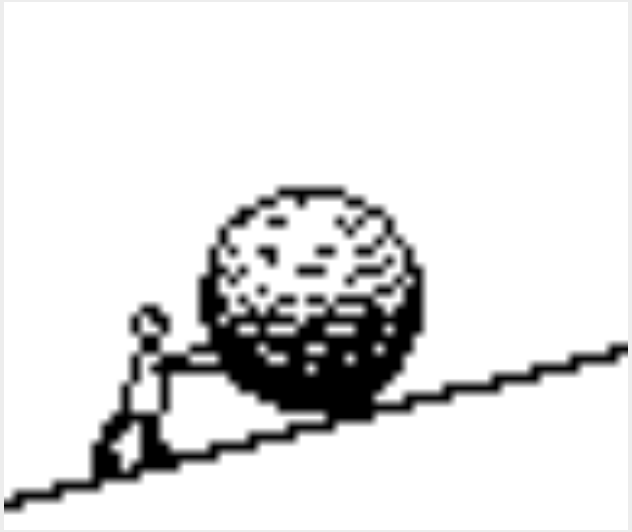
- Information and knowledge processing, including natural language processing, information extraction, integration of data from heterogeneous sources or domains, event detection, feature recognition.
- Tools for analyzing and/or storing very large datasets, data supporting ongoing experiments, and other data used in scientific research.
- Knowledge representation, including vocabularies, ontologies, simulations, and virtual reality.
- Linkage of experimental and model results to benefit research.
- Innovative uses of information technology in science applications, including decision support, error reduction, outcomes analysis, and information at the point of end-use.
- Efficient management and utilization of information and data, including knowledge acquisition and management, process modeling, data mining, acquisition and dissemination, novel visual presentations, and stewardship of large-scale data repositories and archives.
- Human-machine interaction, including interface design, use and understanding of science discipline-specific information, information needs, and uses.
- High-performance computing and communications relating to scientific applications, including efficient machine-machine interfaces, transmission and storage, real-time decision support.
- Innovative uses of information technology to enhance learning, retention and understanding of science discipline-specific information.
- REFERENCE: http://grants.nih.gov/grants/guide/pa-files/PA-06-094.html

General Themes in Informatics Research

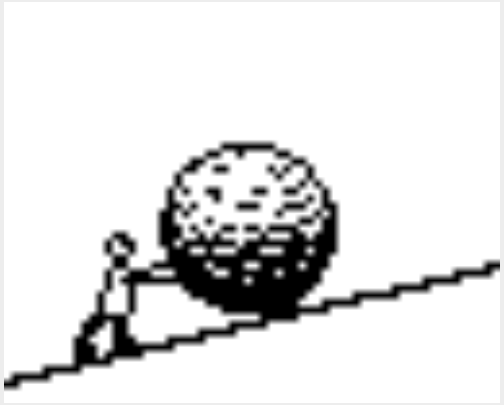
- Information and knowledge processing, including natural language processing, information extraction, integration of data from heterogeneous sources or domains, event detection, feature recognition.
- Tools for analyzing and/or storing very large datasets, data supporting ongoing experiments, and other data used in scientific research.
- Knowledge representation, including vocabularies, ontologies, simulations, and virtual reality.
- Linkage of experimental and model results to benefit research.
- Innovative uses of information technology in science applications, including decision support, error reduction, outcomes analysis, and information at the point of end-use.
- Efficient management and utilization of information and data, including knowledge acquisition and management, process modeling, data mining, acquisition and dissemination, novel visual presentations, and stewardship of large-scale data repositories and archives.
- **Human-machine interaction, including interface design, use and understanding of science discipline-specific information, information needs, and uses.**
- High-performance computing and communications relating to scientific applications, including efficient machine-machine interfaces, transmission and storage, real-time decision support.
- **Innovative uses of information technology to enhance learning, retention and understanding of science discipline-specific information.**
- REFERENCE: http://grants.nih.gov/grants/guide/pa-files/PA-06-094.html

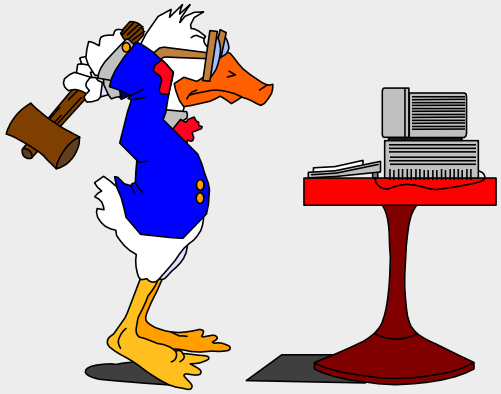
How will you handle too much data and too much information?

The old technology might just run off the tracks! ...or it might run us over!



How will we respond ?





We need something better …

We need something better, Jim !

We need computers …
but not the usual kind !

We need the classical kind
(which pre-dates computing devices)

Modes of Computing

- **Numerical Computation (*in silico*)**
  - Fast, efficient
  - Processing power is rapidly increasing
  - Model-dependent, subjective, only as good as your best hypothesis
- **Computational Intelligence**
  - Data-driven, objective (machine learning)
  - Often relies on human-generated training data
  - Often generated by a single investigator
  - Primitive algorithms
  - Not as good as humans on most tasks
- **Human Computation (*Carbon-based Computing*)**
  - Data-driven, objective (human cognition)
  - Creates training sets, Cross-checks machine results
  - Excellent at finding patterns, image classification
  - Capable of classifying anomalies that machines don't understand
  - Slow at numerical processing, low bandwidth, easily distracted

Outline

- The Emergence of Informatics Research in the Sciences
- **Human Computation**
- Citizen Science: Introduction
- Citizen Science: Examples
- The Zooniverse Project
- The Zooniverse Machine Learning Challenges

What is Human Computation?

- Luis von Ahn says ... it is ...
  - Computation that is “impossible” by computers alone:
    - Because there is too much data, or
    - Problem is essentially impossible for a computer. For example:

(captcha)

    - What else do we call this? – ***IMAGE ANNOTATION***:
      - The ESP Game (Google Images, flickr); or Astronomical Image or Event annotation when # of images or # of events is very large!

The ESP Game: http://www.espgame.org/

User #1 sees this screen, adding his/her own choices of words to describe the image.

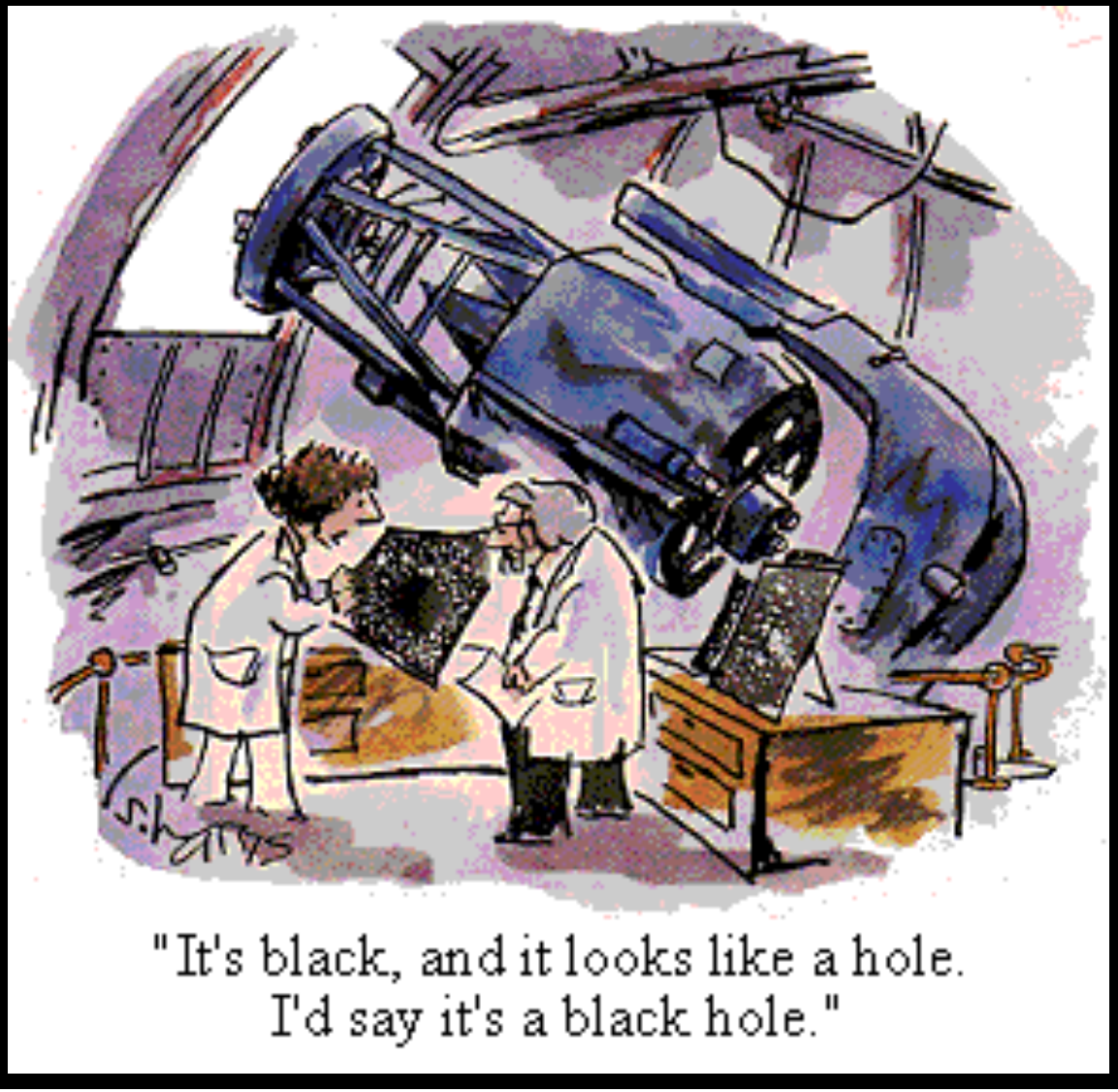
User #2 sees the same image, but on their own screen, adding their own word choices.

Both players get points when their words agree.

It takes a human to interpret a complex image

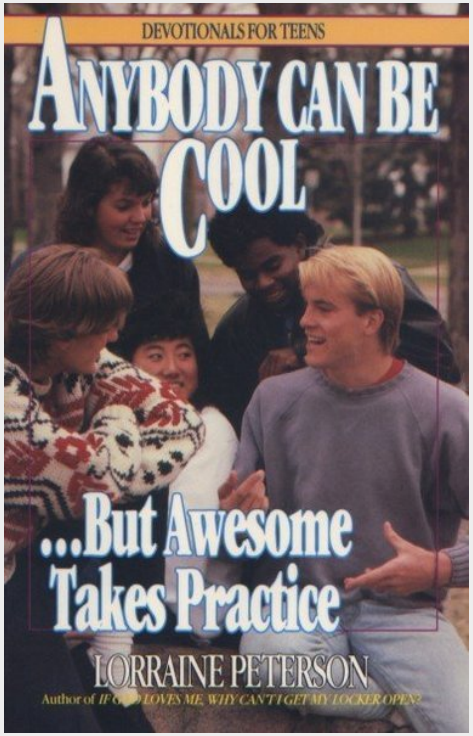
It takes a human to interpret a complex image

usually ...



The emergence of citizen science!

Anybody can participate in the science discovery process



Outline

- The Emergence of Informatics Research in the Sciences
- Human Computation
- **Citizen Science: Introduction**
- Citizen Science: Examples
- The Zooniverse Project
- The Zooniverse Machine Learning Challenges

Science@home vs. Citizen Science

- The BOINC projects (SETI@home, LHC@home, Climateprediction.net, Folding@home, Einstein@home) use your idle computer time
- Those projects do not require you to do anything more than download the screensaver software for your computer.
- BOINC wakes up when you and your computer go to sleep.

Citizen Science

- Exploits the cognitive abilities of **Human Computation**!
- Novel mode of data collection:
  - Citizen Science! = Volunteer Science = Participatory Science
  - *e.g.,* VGI = Volunteer Geographic Information (Goodchild '07)
  - *e.g.,* Galaxy Zoo @ http://www.galaxyzoo.org/
- Citizen science refers to the involvement of volunteer non-professionals in the research enterprise.
- The Citizen Science experience …
  - must be engaging,
  - must work with real scientific data/information (all of it),
  - must not be busy-work (all clicks must count),
  - **must address authentic science research questions** that are beyond the capacity of science teams and enterprises, and
  - must involve the scientists.

Outline

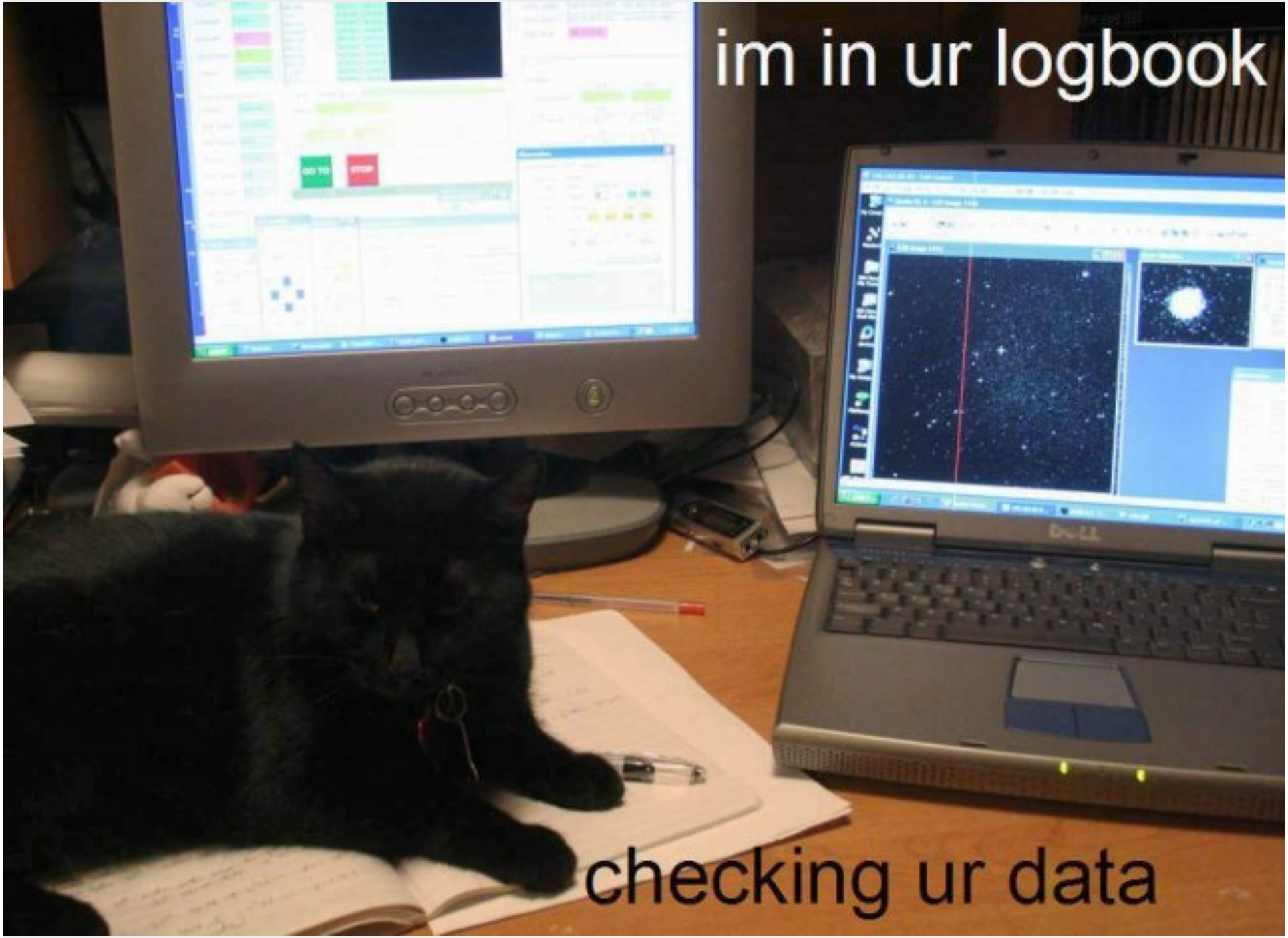
- The Emergence of Informatics Research in the Sciences
- Human Computation
- Citizen Science: Introduction
- **Citizen Science: Examples**
- The Zooniverse Project
- The Zooniverse Machine Learning Challenges

Examples of Volunteer Science

- AAVSO (Amer. Assoc. of Variable Star Observers)
- Audubon Bird Counts
- Project Budburst
- Stardust@Home
- VGI (Volunteer Geographic Information)
- CoCoRaHS (Community Collaborative Rain, Hail and Snow network)
- Galaxy Zoo **(~20 refereed pubs so far…)**
- Zooniverse (buffet of Zoos)
- U-Science (semantic science 2.0) [ref: Borne 2009]
  - includes Biodas.org, Wikiproteins, HPKB, AstroDAS
  - **U**biquitous, **U**ser-oriented, **U**ser-led, **U**niversal, **U**ntethered, **You**-centric **Science**

Anybody can participate and contribute to the science...

"On the Internet, nobody knows you're a dog"

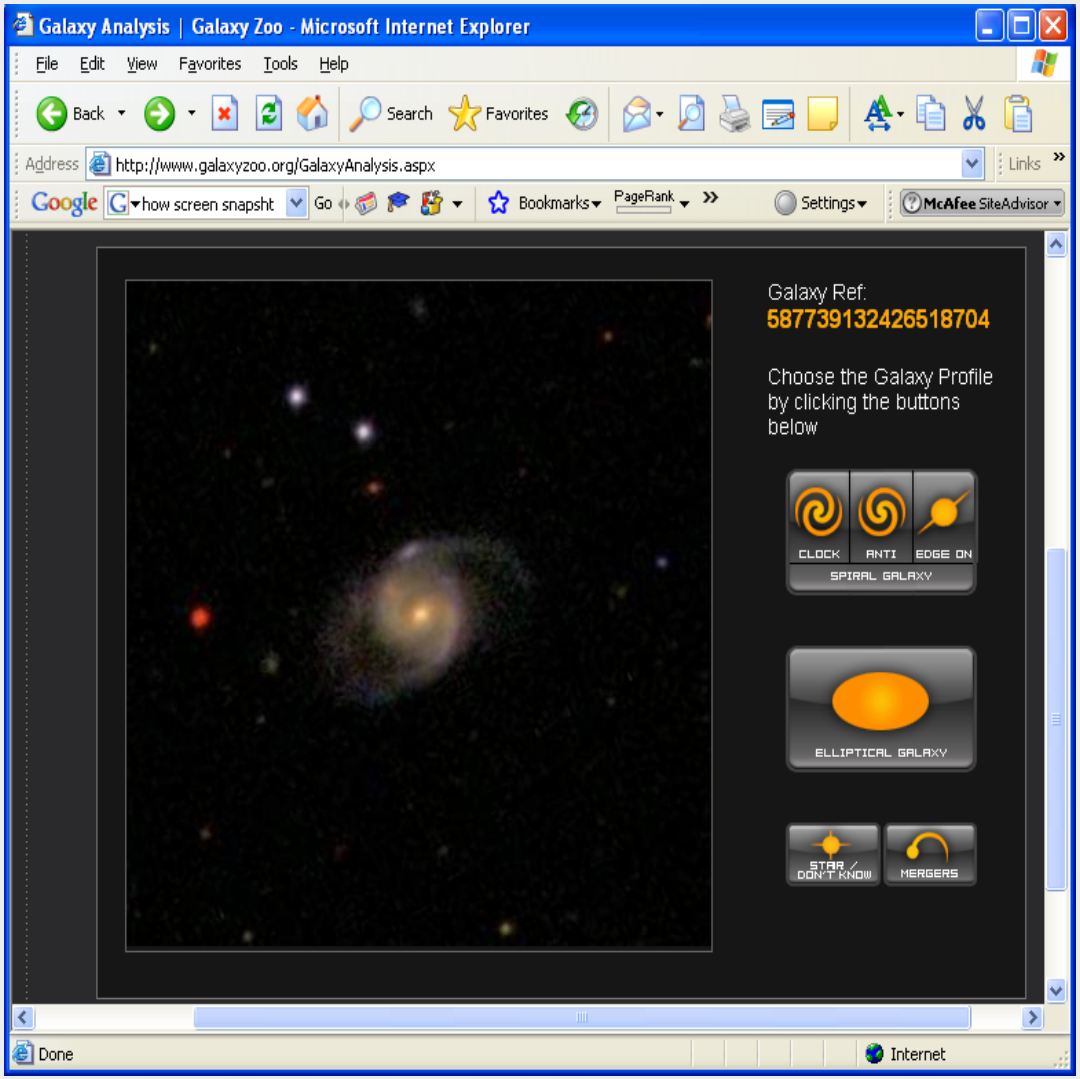

im in ur logbook
checking ur data

GalaxyZoo: http://www.galaxyzoo.org/

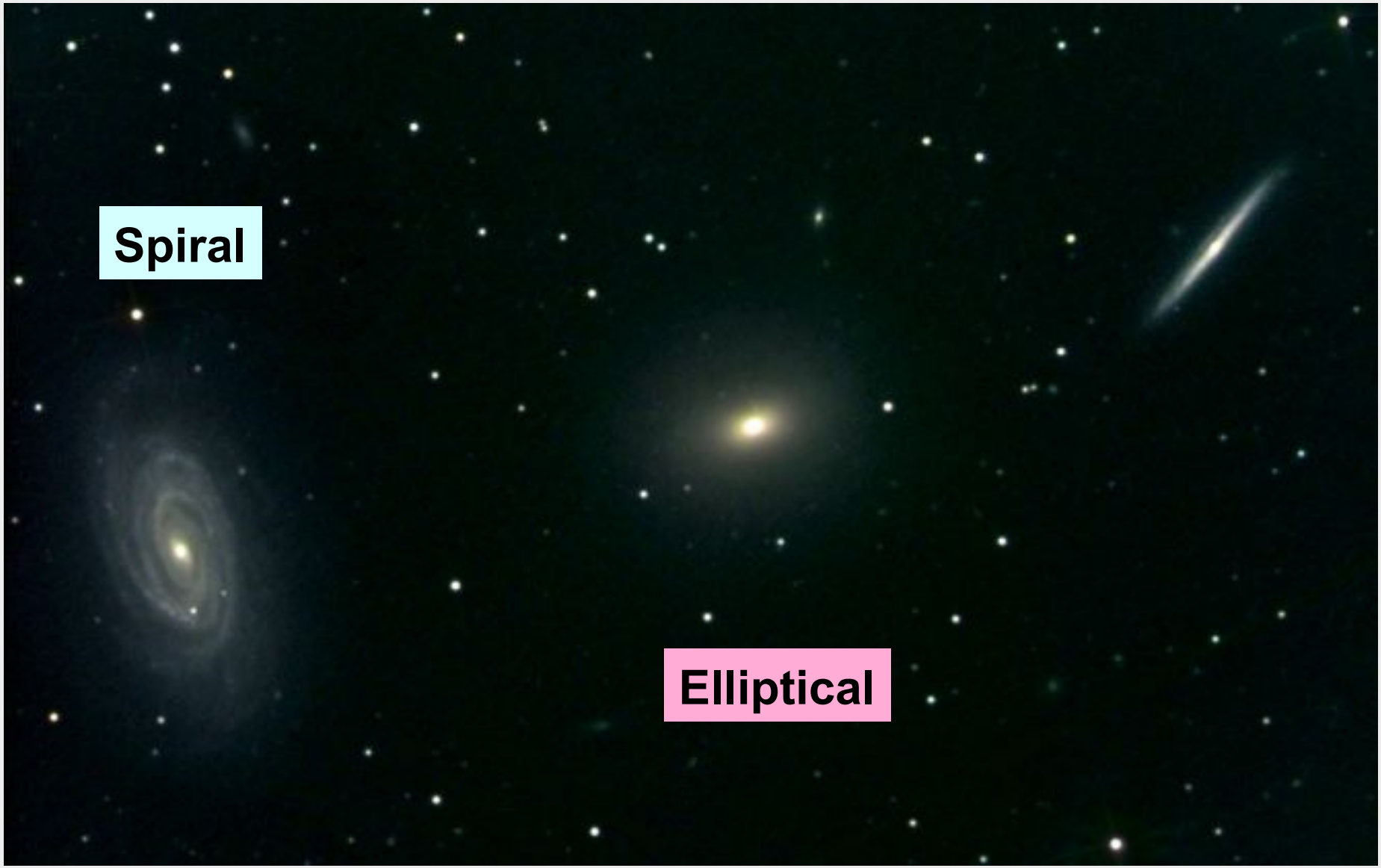
You can help us to classify a million galaxies!

- **"Welcome to GalaxyZoo**, the project which harnesses the power of the internet - and your brain - to classify a million galaxies. By taking part, you'll not only be contributing to scientific research, but you'll view parts of the Universe that literally no-one has ever seen before and get a sense of the glorious diversity of galaxies that pepper the sky."

- **"Why do we need you?** – The simple answer is that the human brain is much better at recognizing patterns than a computer can ever be. Any computer program we write to sort our galaxies into categories would do a reasonable job, but it would also inevitably throw out the unusual, the weird and the wonderful. To rescue these interesting systems which have a story to tell, we need you."



There are 2 main types of galaxies: Spiral & Elliptical (plus there are some peculiar & irregular galaxies)

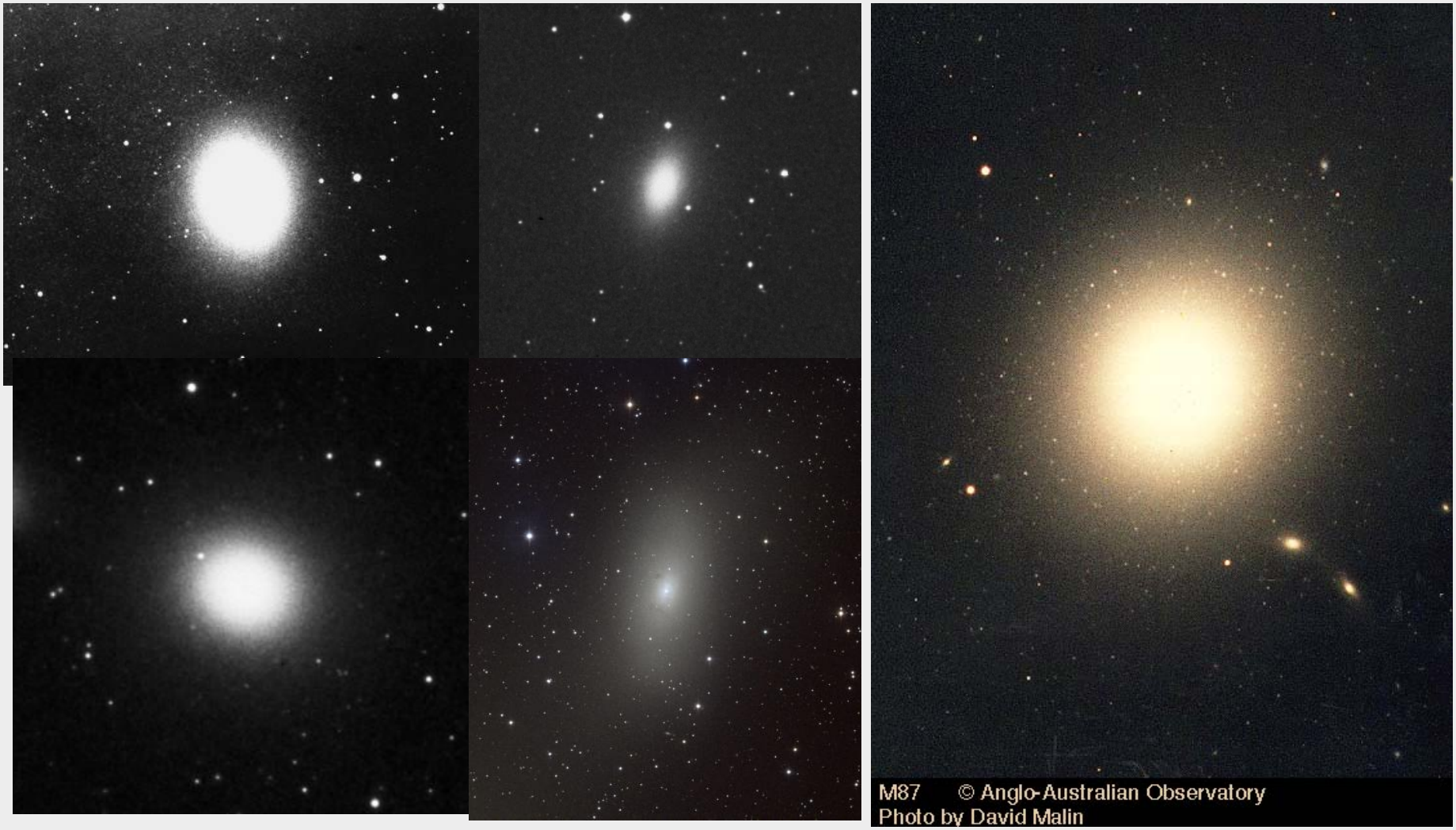


Gallery of Elliptical Galaxies

M32

M59

M87

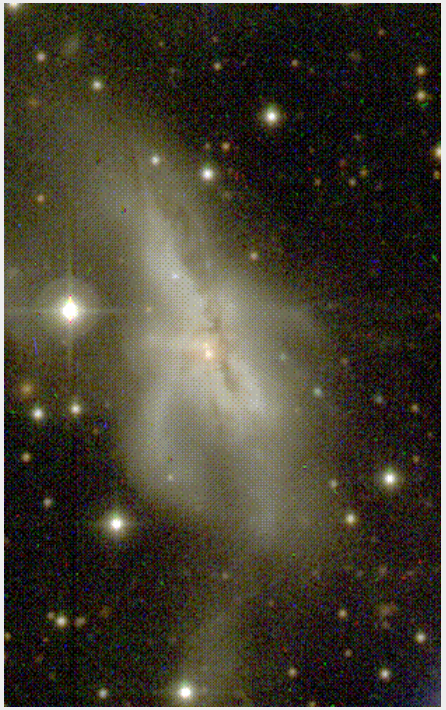


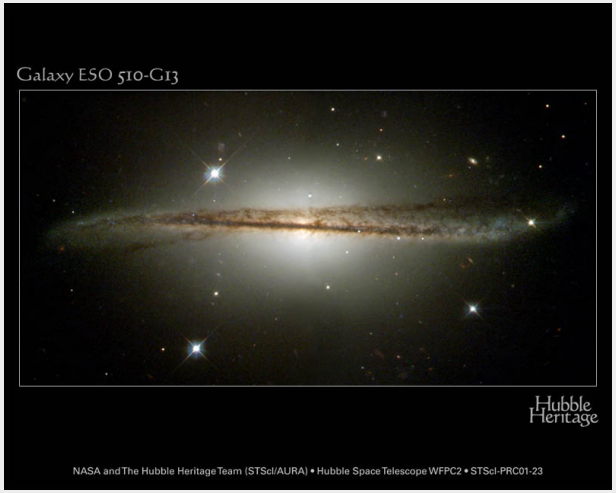
M105

M110=NGC205

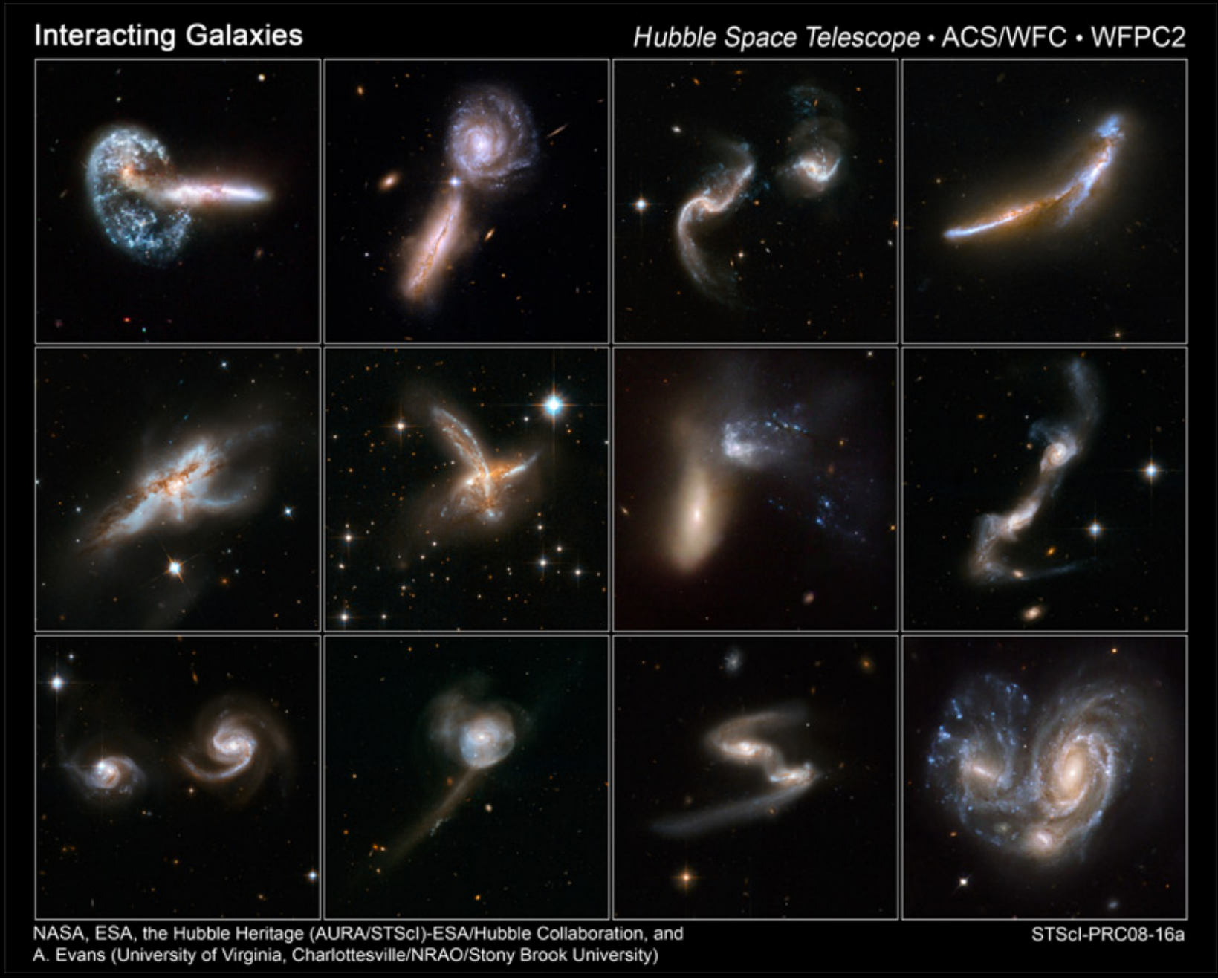
Gallery of Face-on Spiral Galaxies: studying their properties indicates that our Milky Way is a Spiral

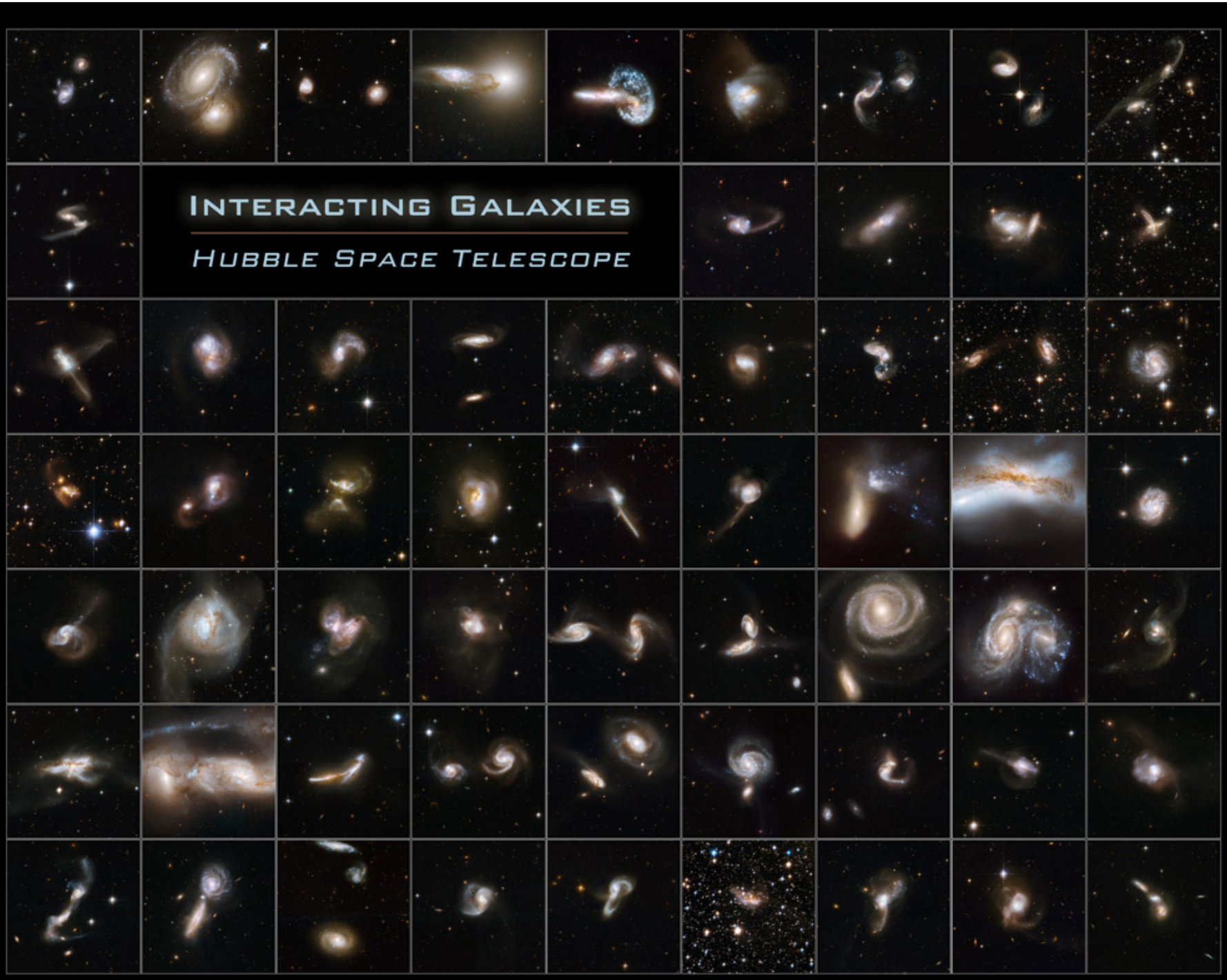
There are lots of Peculiar Galaxies also !





Galaxies Gone Wild !




Interacting Galaxies
Hubble Space Telescope

Merging/Colliding Galaxies are the building blocks of the Universe: 1+1=1

More gorgeous Colliding Galaxies !

NGC 6050

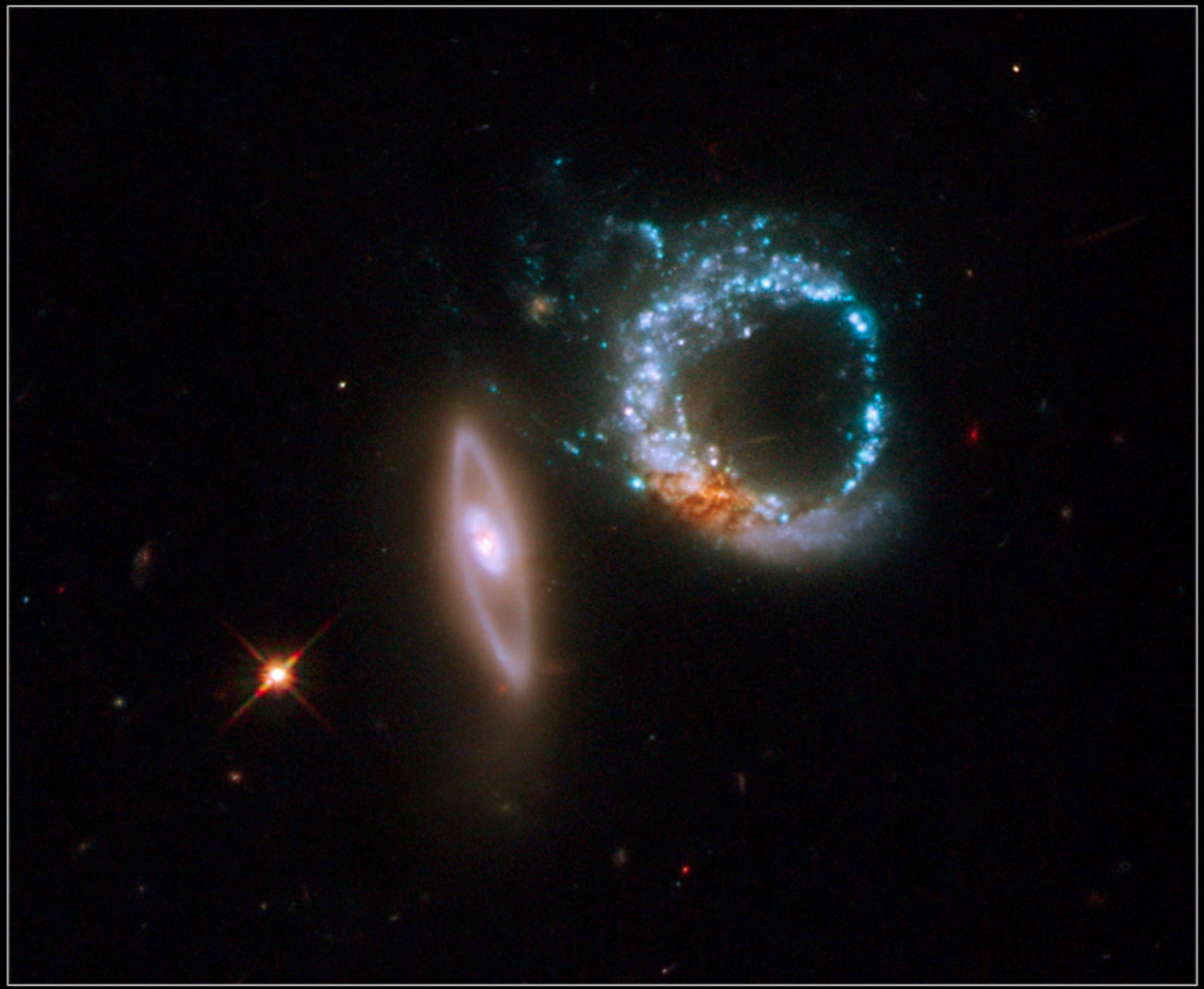
Arp 148

Hubble Heritage

NASA, ESA, and The Hubble Heritage (AURA/STScI)-ESA/Hubble Collaboration • *Hubble Space Telescope* WFPC2

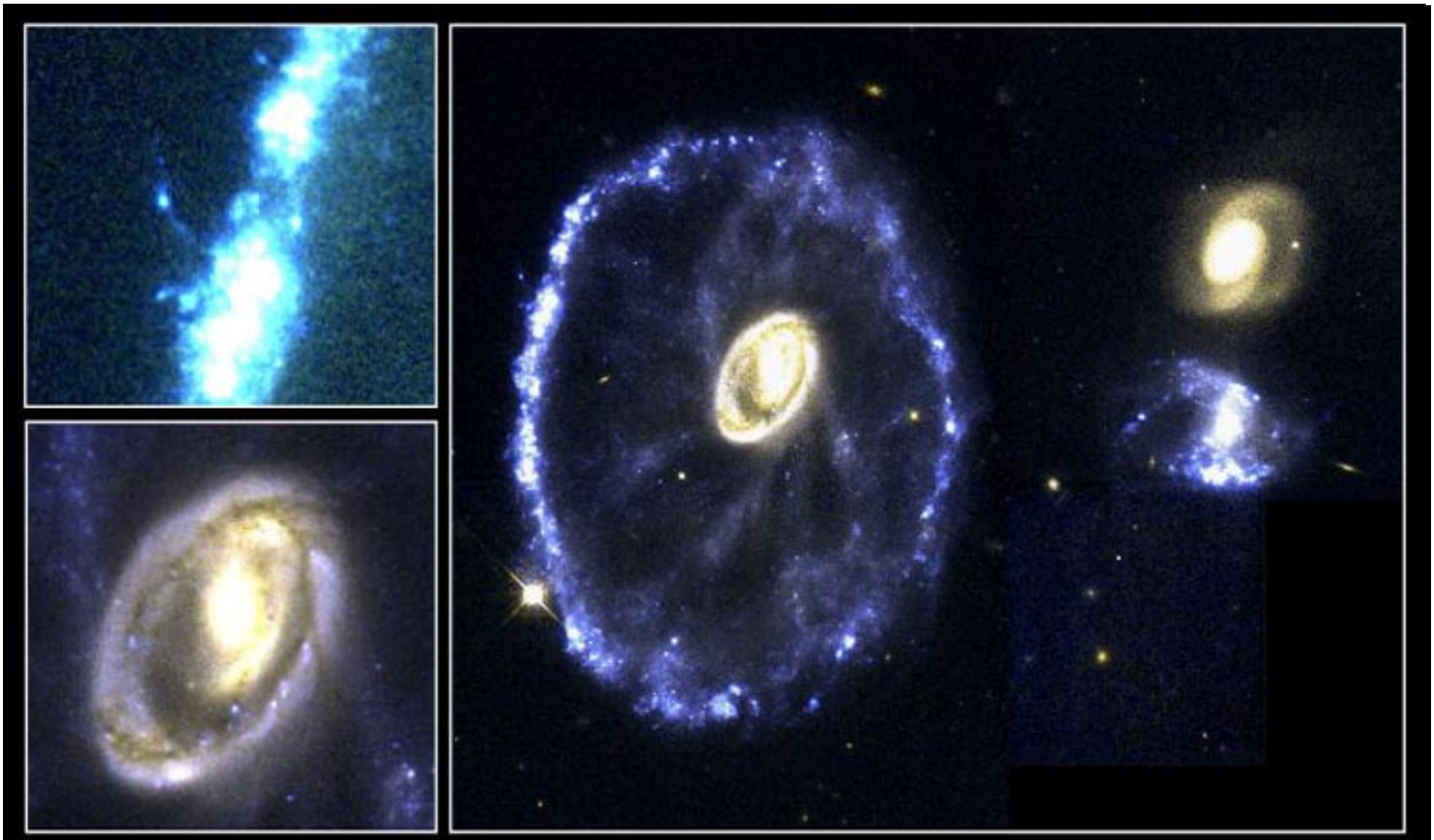
Interacting Galaxies Arp 147

Hubble Space Telescope • WFPC2



NASA, ESA, and M. Livio (STScI)

STScI-PRC08-37



http://hubblesite.org/gallery/album/pr1995002a


The Universe is full of galaxies

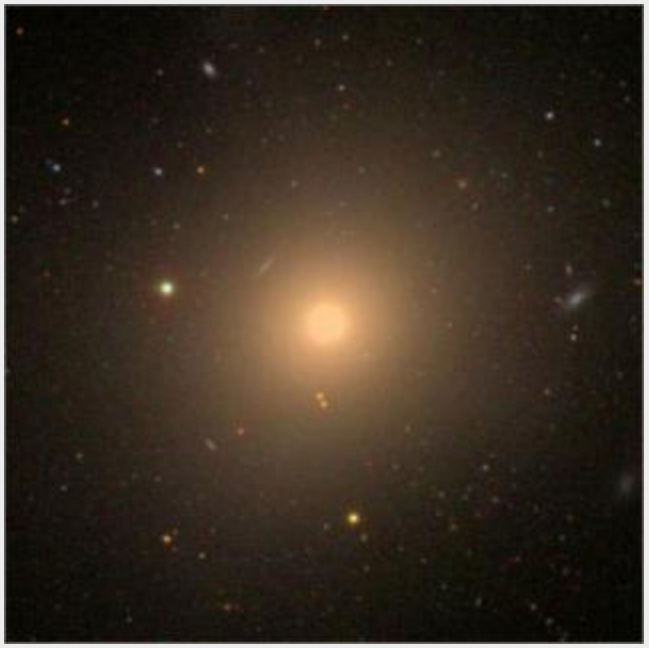
Astronomers have collected images of hundreds of millions of galaxies, but we have analyzed maybe only 10% of all of these!

In the next 10-20 years, we will have images of tens of billions of new galaxies!!

How will we identify and classify all of these spirals, ellipticals, and mergers?

Spirals, ellipticals, and mergers? Oh my!

Galaxy Zoo helps scientists by engaging the public (hundreds of thousands of us) to classify millions of galaxies:
Is it a Spiral Galaxy or Elliptical Galaxy?



- Galaxy Zoo project:
  - ~390,000 participants (*and growing*)
  - ~1 million galaxies have been labeled (classified)
  - ~180 million classifications have been collected

GALAXY ZOO 2

Home | How To Take Part | My Galaxies | Contact Us | Profile | Logout

INVERT GALAXY IMAGE | ADD TO MY FAVOURITES

Classify Galaxies

Answer the question below using the buttons provided.

Is the galaxy simply smooth and rounded, with no sign of a disk?


Smooth


Features or disk


Star or artifact

Need help?

GALAXY ZOO 2

Home | How To Take Part | My Galaxies | Contact Us | Profile | Logout

INVERT GALAXY IMAGE | ADD TO MY FAVOURITES

Classify Galaxies

Answer the question below using the buttons provided.

Is the galaxy simply smooth and rounded, with no sign of a disk?

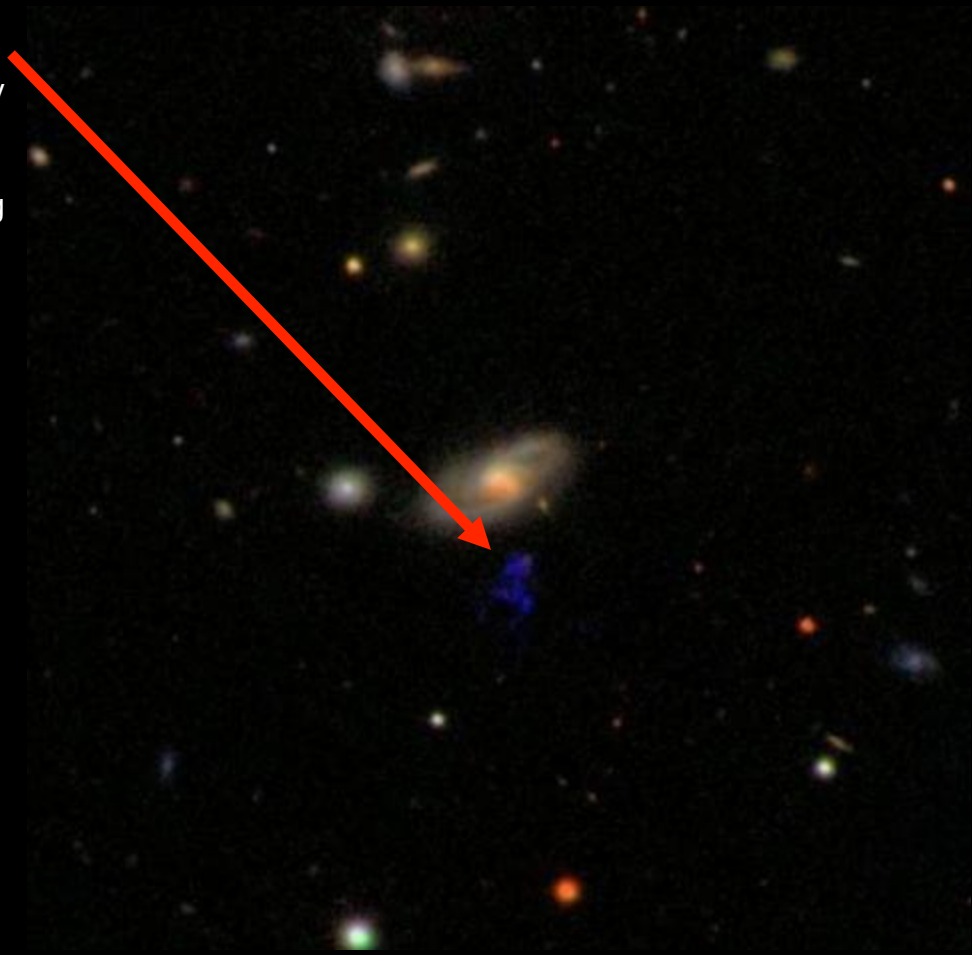
Smooth | Features or disk | Star or artifact

Need help?


Help scientists to find Merging/Colliding Galaxies:
they are neither spiral nor elliptical

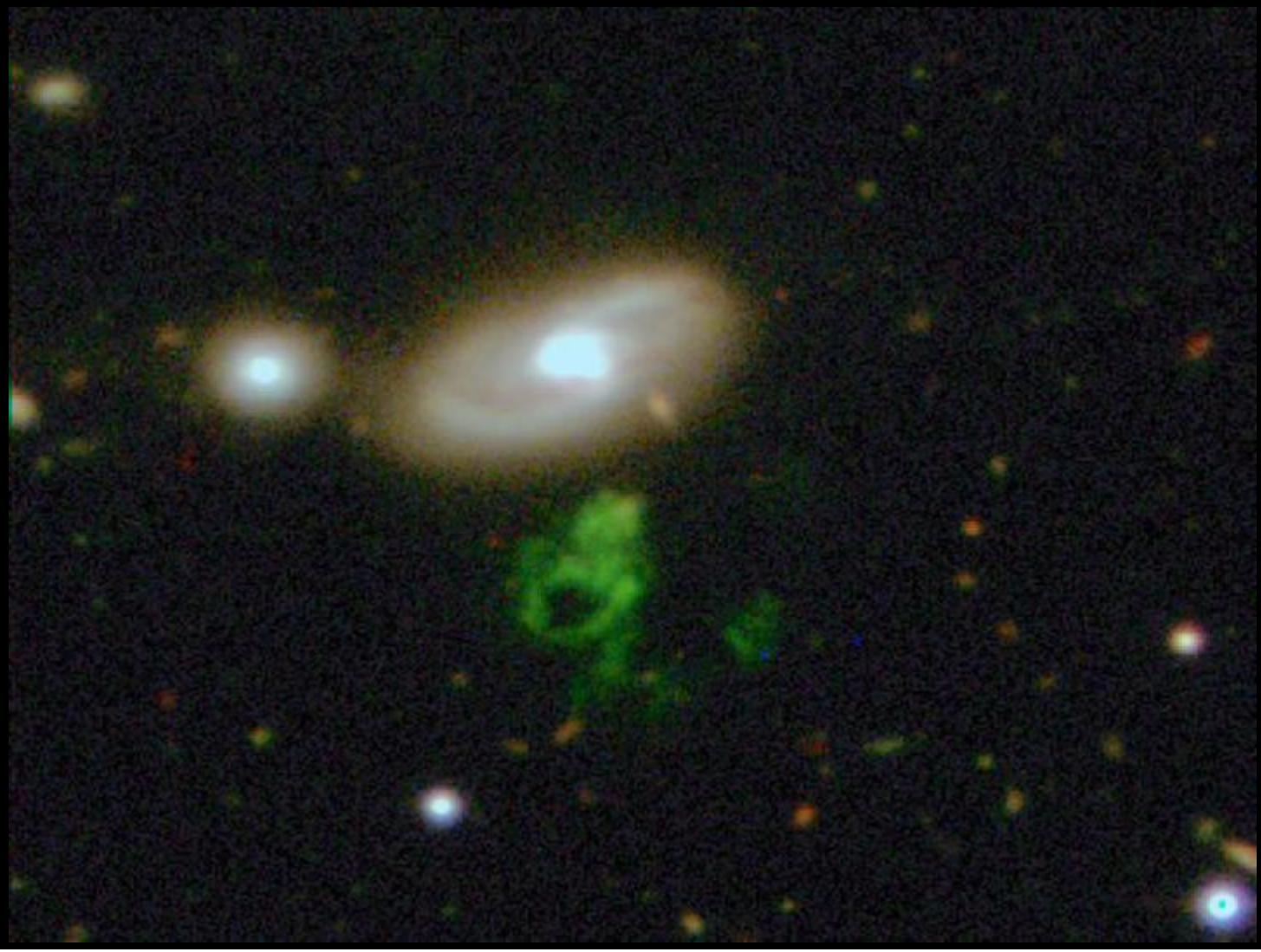
Hanny's Voorwerp: Hanny's Object – found by a school teacher in Holland

- **What is that greenish blue thing?**
- A volunteer sky enthusiast surfing through online Galaxy Zoo images has discovered something really strange. The mystery object is unusually green, not of any clear galaxy type, and situated below relatively normal looking spiral galaxy IC 2497. Dutch schoolteacher Hanny van Arkel, discovered the strange green "voorwerp" (Dutch for "object") last year. The Galaxy Zoo project encourages sky enthusiasts to browse through SDSS images and classify galaxy types. Now known popularly as Hanny's Voorwerp, subsequent observations have shown that the mysterious green blob has the same distance as neighboring galaxy IC 2497. Research is ongoing, but one leading hypothesis holds that Hanny's Voorwerp is a small galaxy that acts like a large reflection nebula, showing the reflected light of a bright quasar event that was visible in the center of IC 2497 about 100,000 years ago. Pictured above, Hanny's Voorwerp was imaged recently by the 2.5-meter Isaac Newton Telescope in the Canary Islands by Dan Smith, Peter Herbert and Chris Lintott (Univ. Hertfordshire). Other collaboration members include Matt Jarvis, Kevin Schawinski, and William Keel.



True color picture of Hanny's Voorwerp:

Hanny's Object – the green blob is probably a light echo from an old Quasar that burned out 100,000 years ago



True color picture of Hanny van Arkel and KB!

Outline

- The Emergence of Informatics Research in the Sciences
- Human Computation
- Citizen Science: Introduction
- Citizen Science: Examples
- **The Zooniverse Project**
- The Zooniverse Machine Learning Challenges

The Zooniverse* :

Advancing Science through User-Guided Learning in Massive Data Streams

* NSF CDI funded program @ http://zooniverse.org

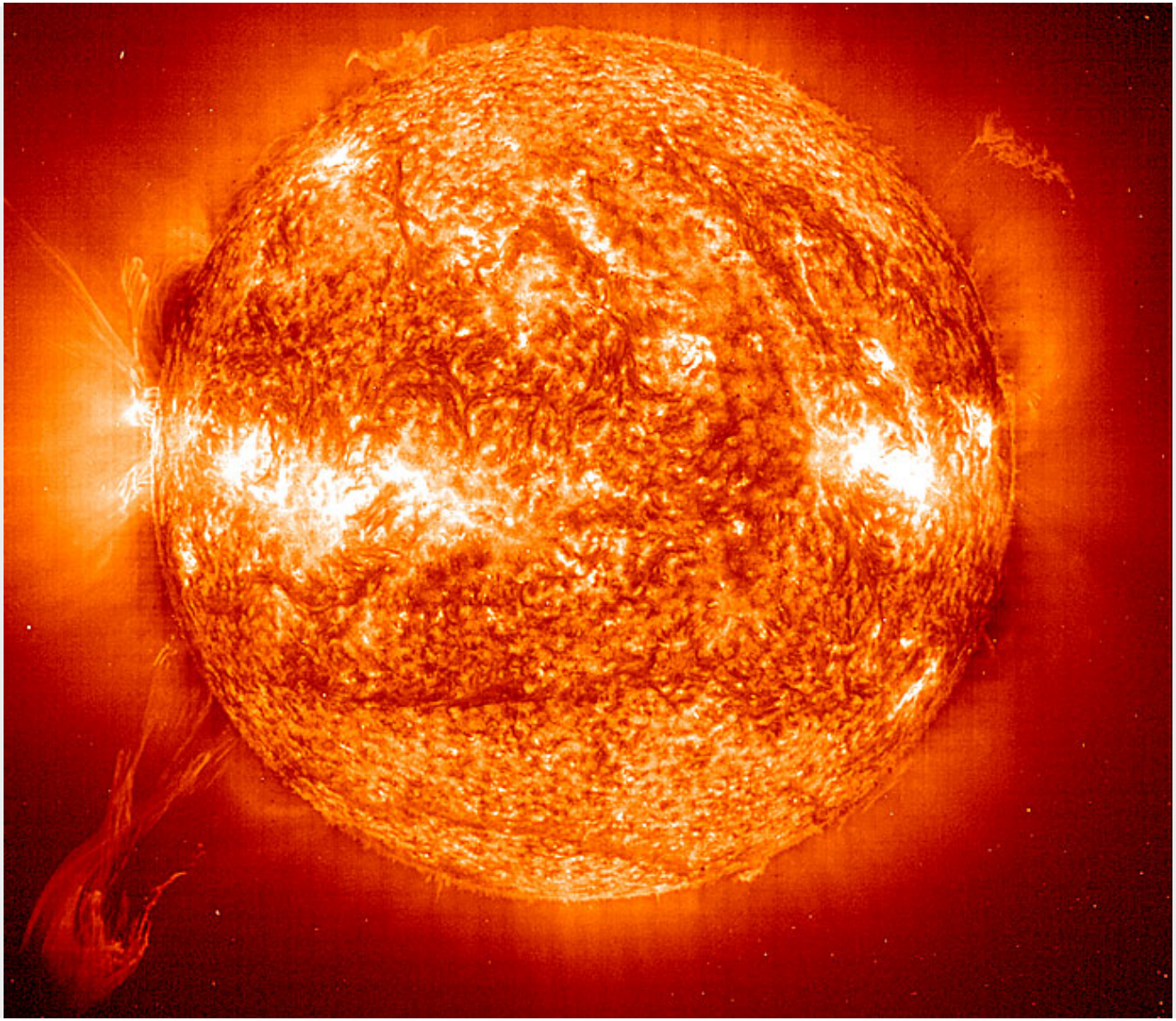
The Zooniverse

http://zooniverse.org/

- New funded NSF CDI grant (PI: L.Fortson [U. Minnesota]; co-PIs J. Wallin [MTSU], K.Borne [GMU], C. Lintott [Oxford])
- Building a framework for new Citizen Science projects, including user-based research tools
- Science domains:
  - Astronomy (Galaxy Merger Zoo, Milky Way Project, Supernova Search, Planet Hunters)
  - The Sun (Solar Storm Watch, with STEREO data)
  - The Moon (Moon Zoo, with LRO data)
  - Egyptology (the Papyri Project)
  - Old Weather (from early 20th century ship logs)
  - And more …

NASA's Solar Dynamics Observatory (SDO):

SDO will generate thousands of images every day, with enormous detail – Citizen Science helps in identifying and sorting out all of the amazing dynamic features on the Sun (CMEs, solar storms).

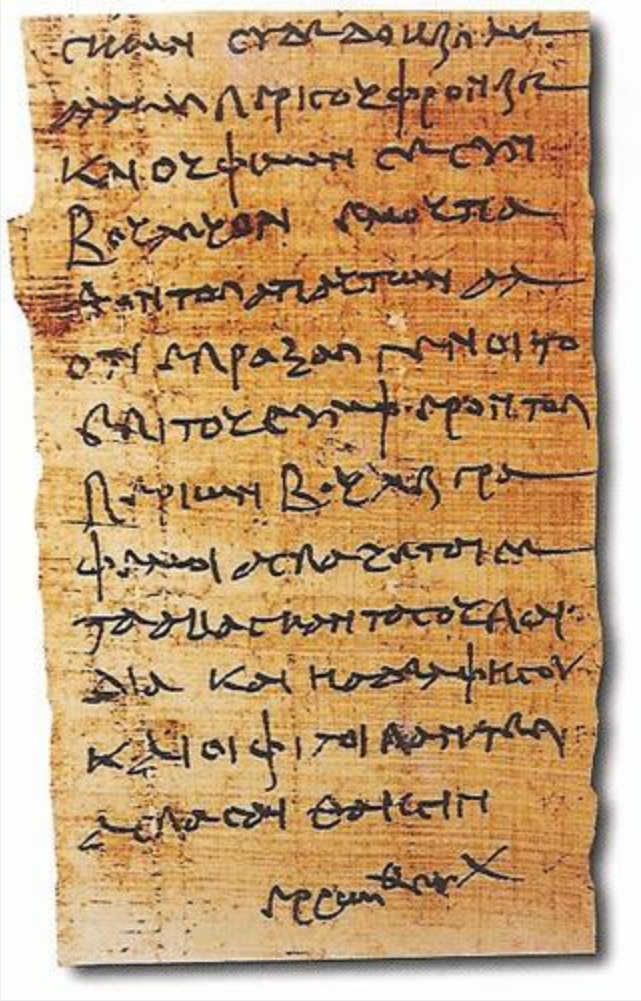


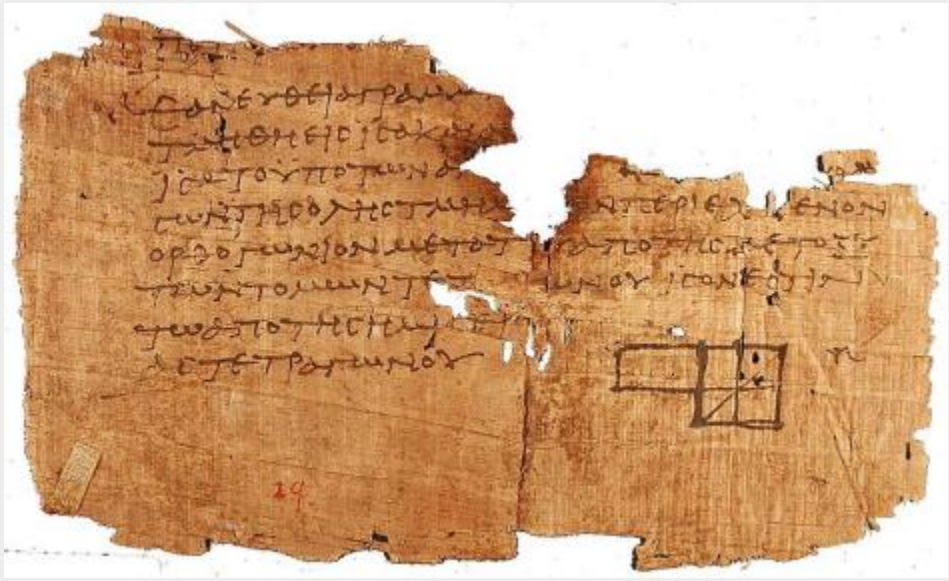
NASA's Lunar Reconnaissance Orbiter:

Citizen Science helps to classify and label all of the different features and regions of the moon (craters, cave lights, free-standing arches)

Egyptology (the Papyri Project)

Oxyrhynchus Papyri Project @ http://www.papyrology.ox.ac.uk/





The Zooniverse: a Buffet of Zoos

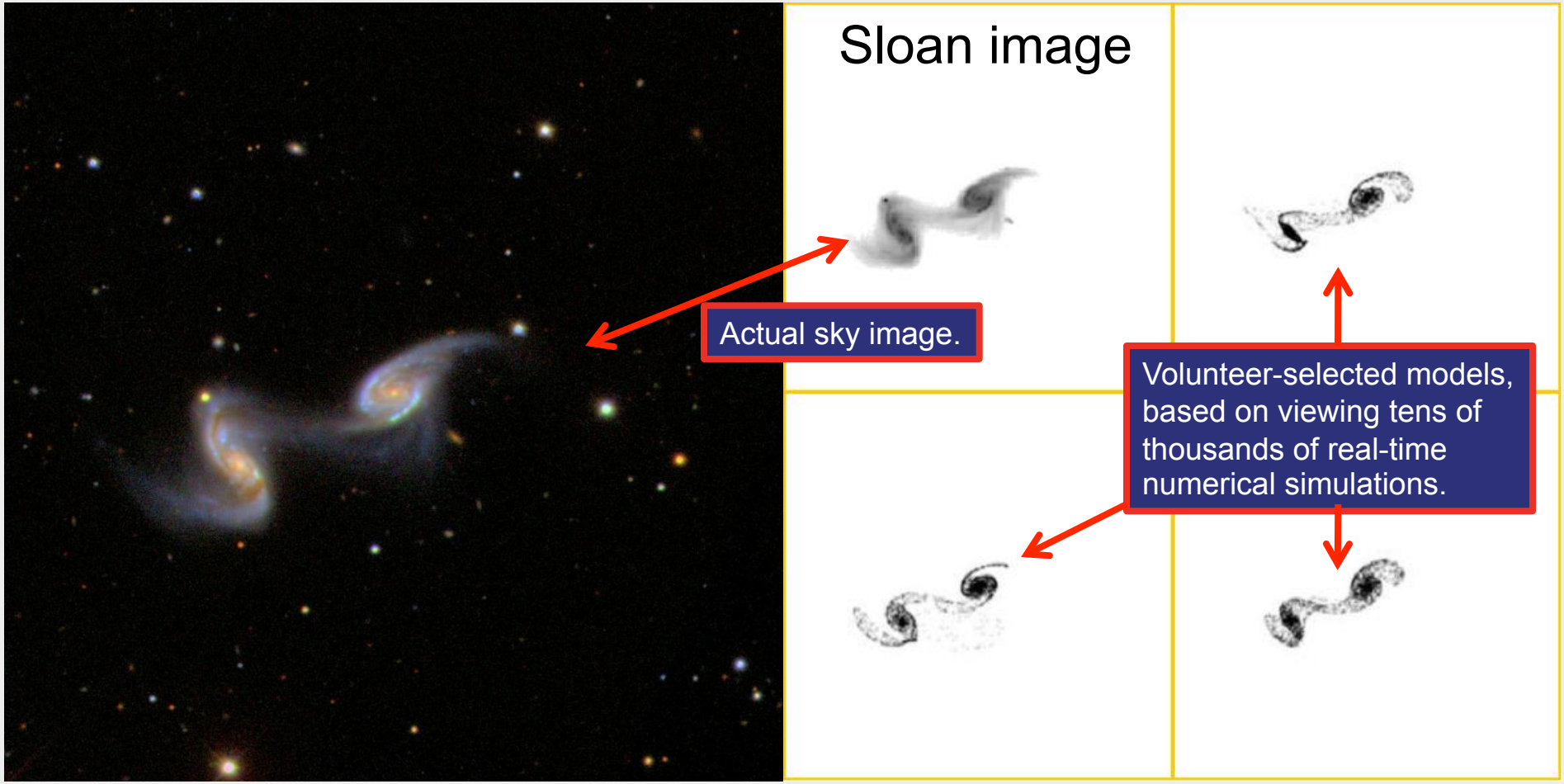
(here are a few examples)

- Galaxy Zoo project (released July 2007): [the original project]
  - http://www.galaxyzoo.org/
    - Classify galaxies (Spiral, Elliptical, Merger, or image artifact)
- **Galaxy Merger Zoo**** (released November 2009)
  - http://mergers.galaxyzoo.org/
    - Run N-body simulations to find best model to match a real merger
    - 18-dimensional model parameter space explored
    - Millions of simulations run, visualized, and tagged by volunteers
- The Hunt for Supernovae (released December 2009)
  - http://supernova.galaxyzoo.org/
    - Real-time event detection and classification
- Moon Zoo (released May 2010)
  - http://www.moonzoo.org/
    - Identify and measure craters and other lunar features in LRO camera data
- The Milky Way Project (released December 2010)
  - http://www.milkywayproject.org/
    - Identify supernova bubbles and other features in infrared map of the Milky Way
    - Excellent description: http://blogs.zooniverse.org/mwp/2011/02/22/reducing-the-data/

**Results presented in the next few slides.

Merging/Colliding Galaxies are the building blocks of the Universe: **1 + 1 = 1**

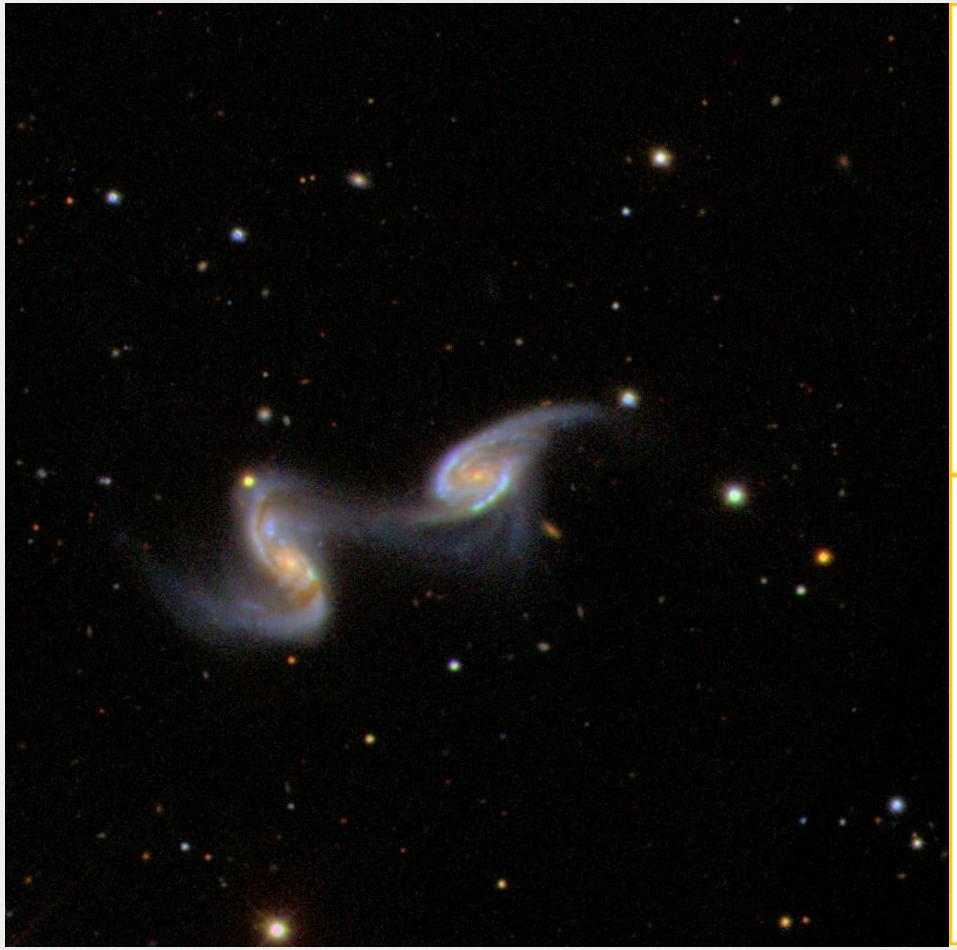
Galaxy Mergers Zoo Gallery

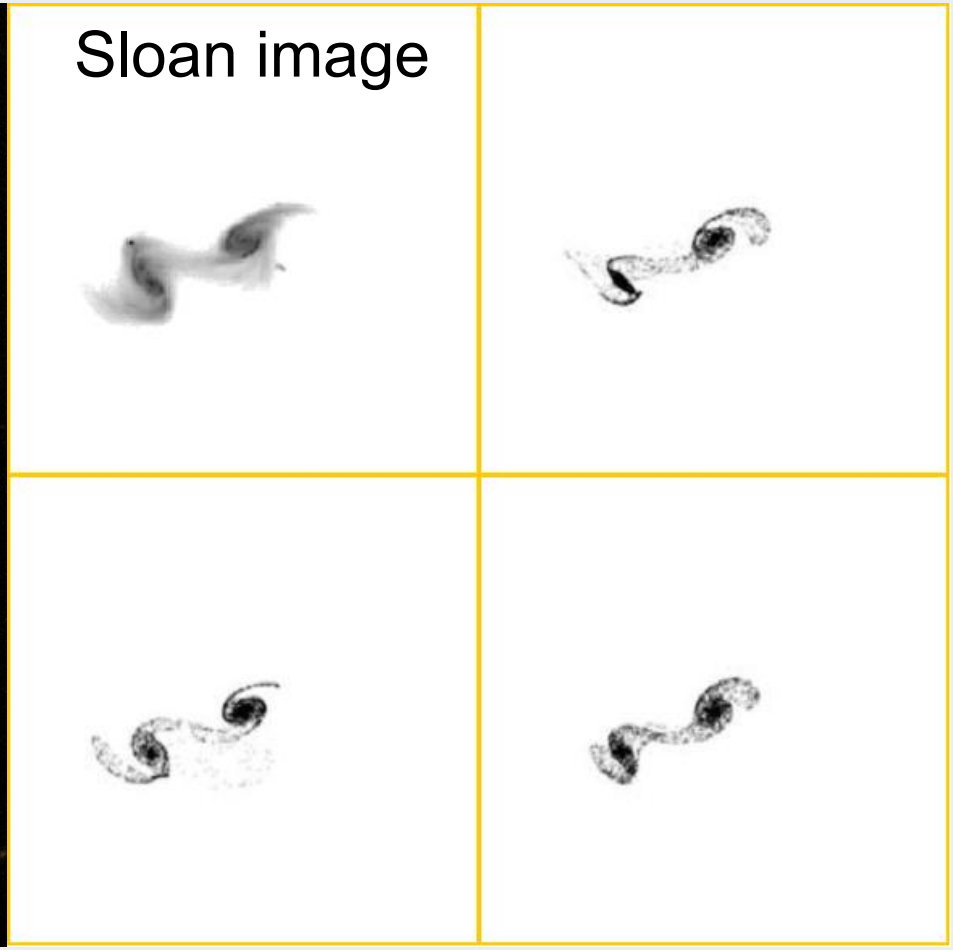


SDSS 587722984435351614

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

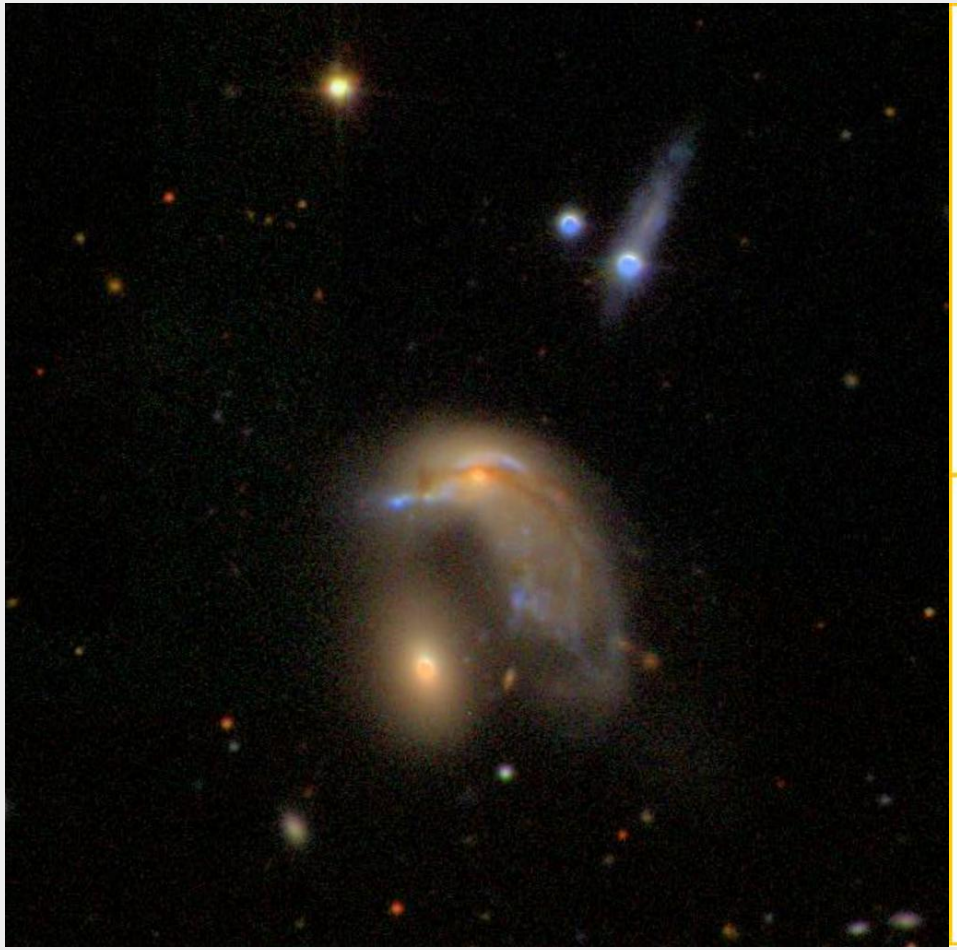
Galaxy Mergers Zoo Gallery





SDSS 587722984435351614

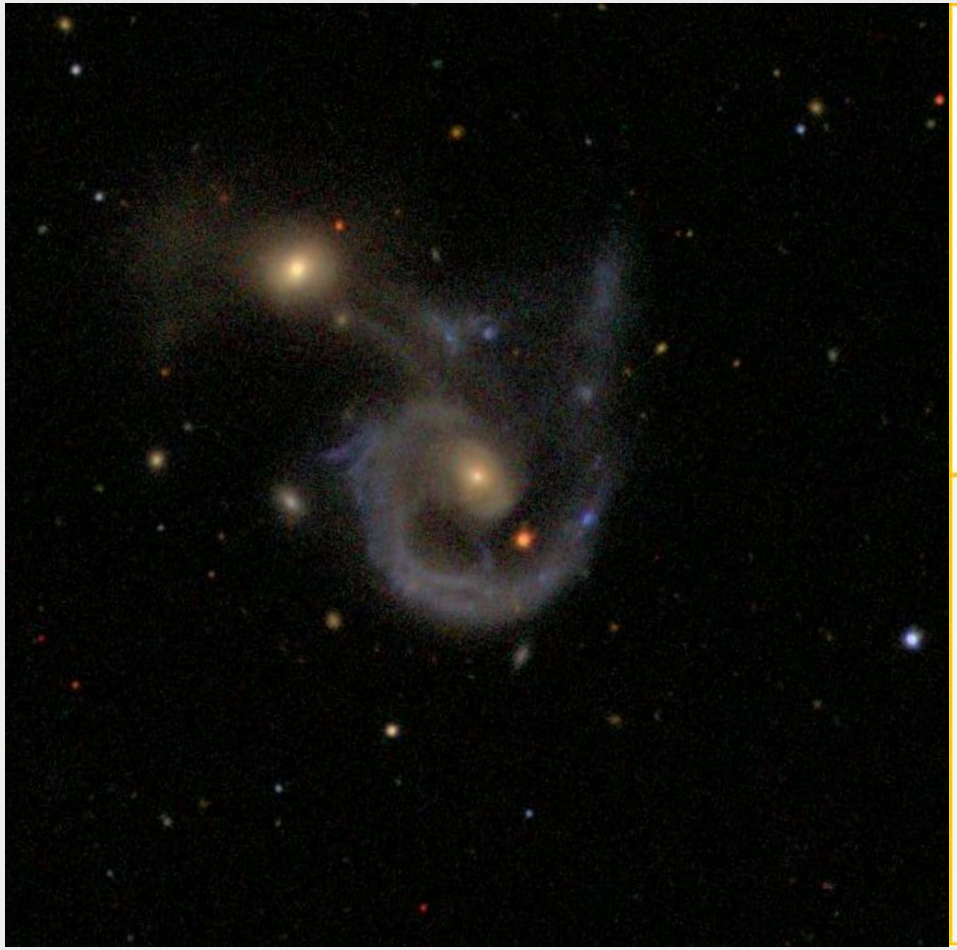
Galaxy Mergers Zoo Gallery

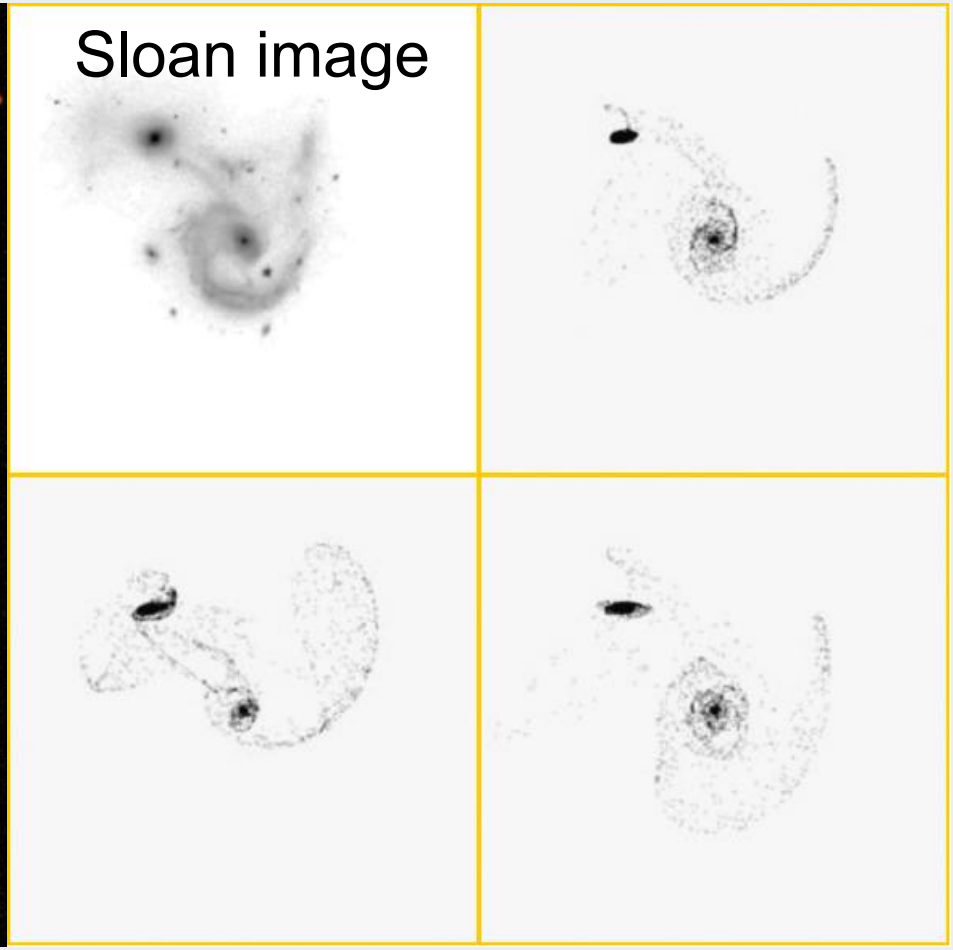


SDSS 587726033843585146

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery

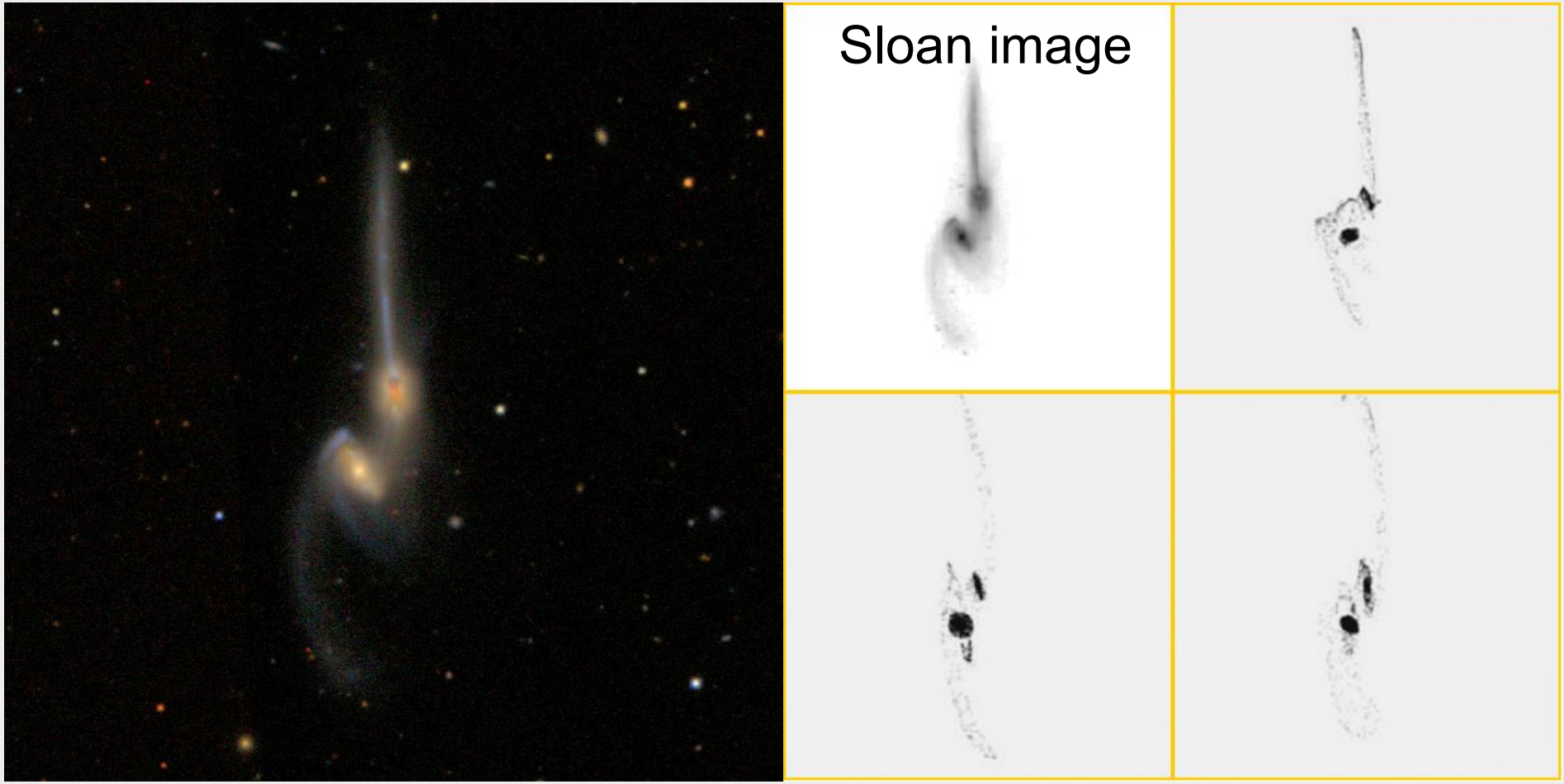




SDSS 587739646743412797

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

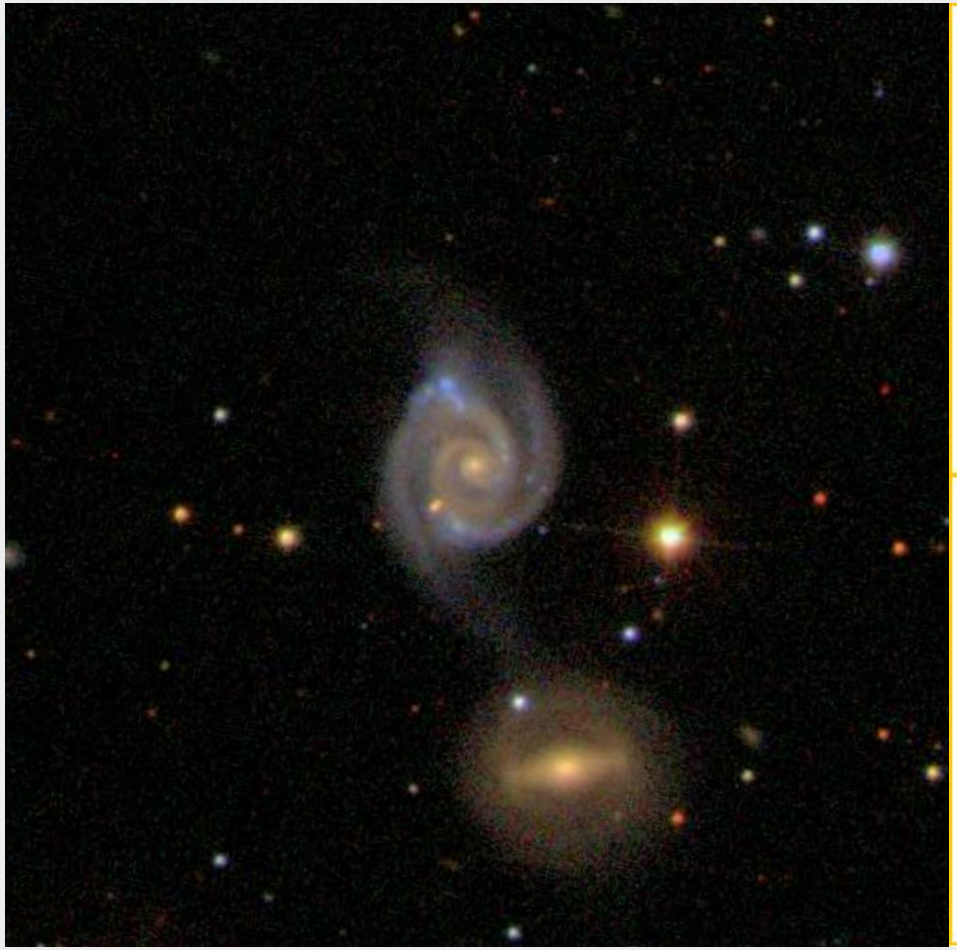
Galaxy Mergers Zoo Gallery

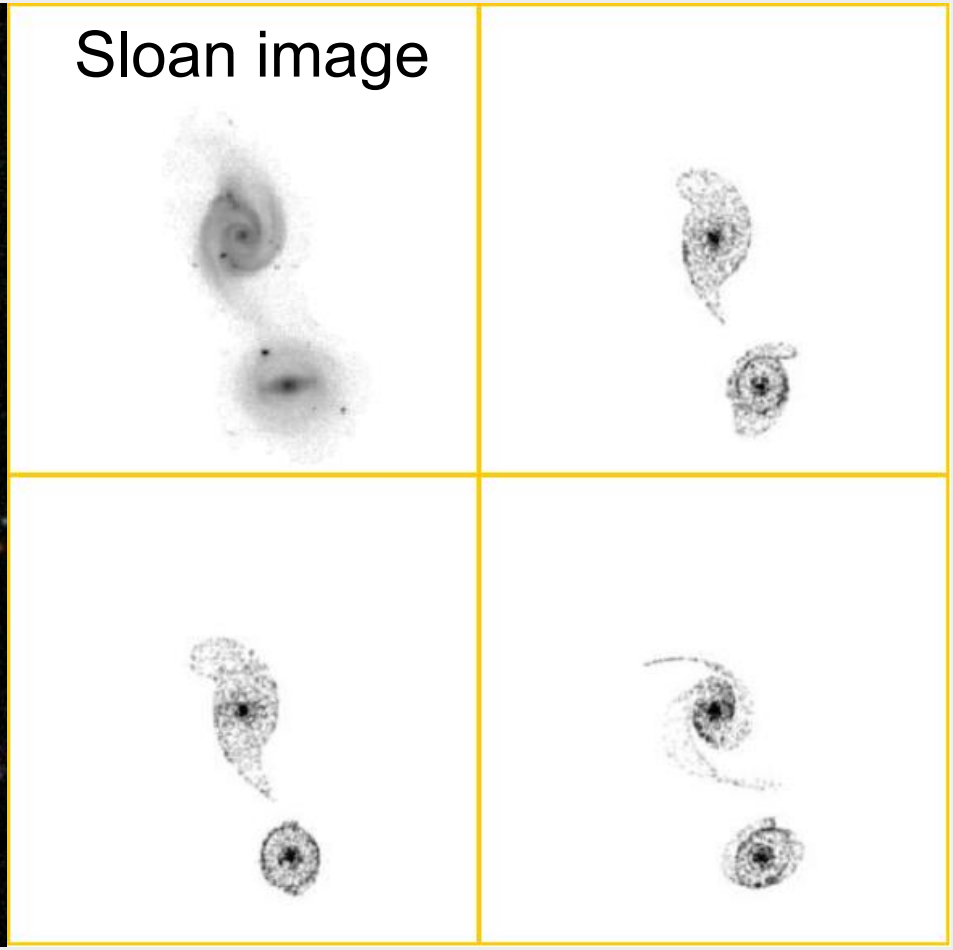


SDSS 587739721900163101

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery

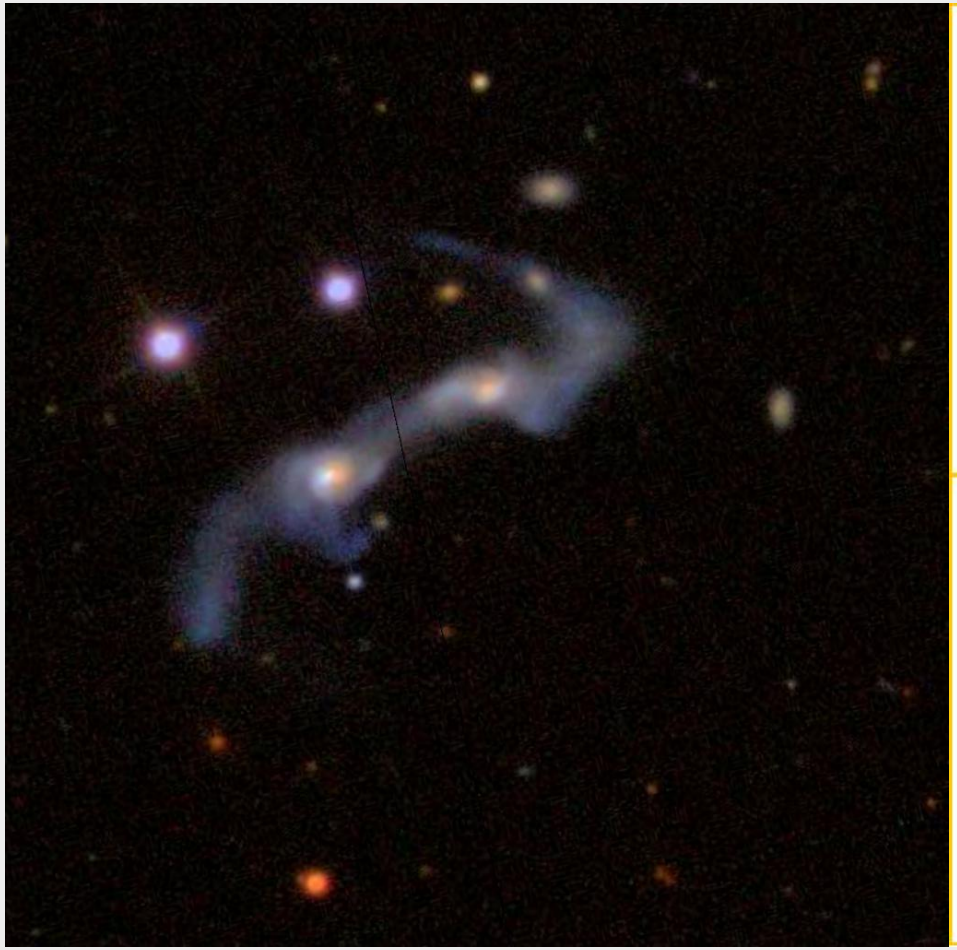


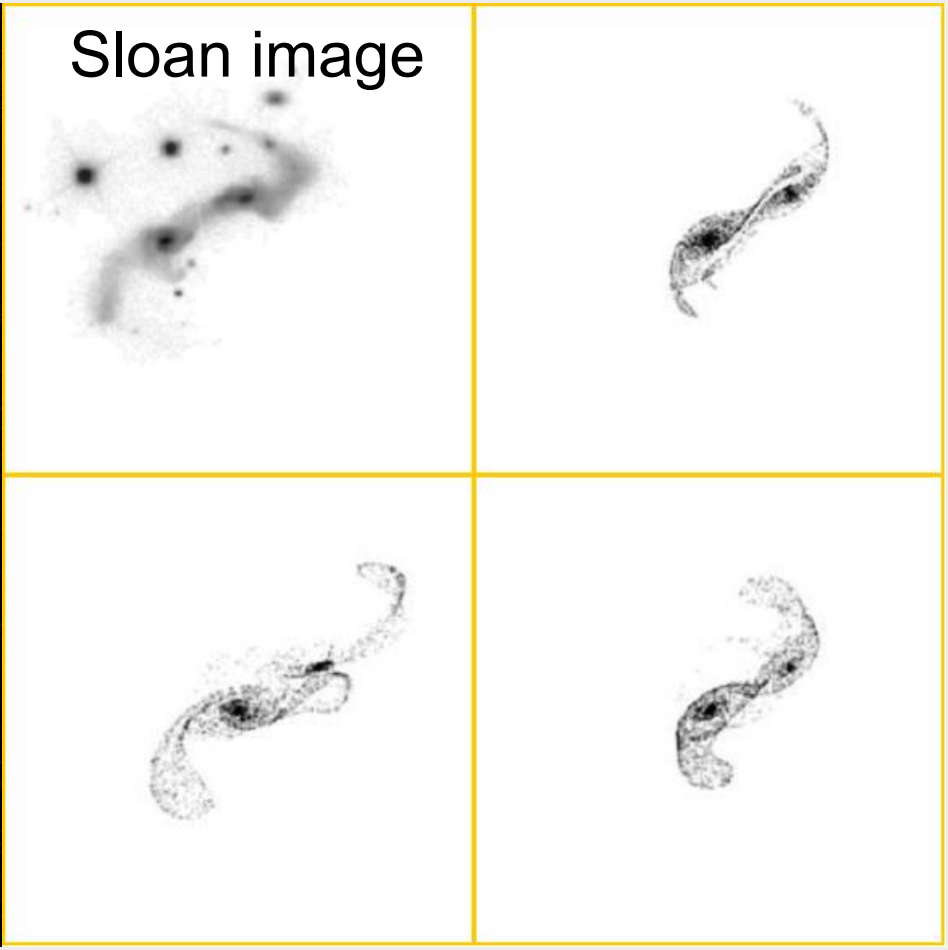


SDSS 587727222471131318

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

Galaxy Mergers Zoo Gallery





SDSS 588011124116422756

GALAXY ZOO
UNDERSTANDING COSMIC MERGERS

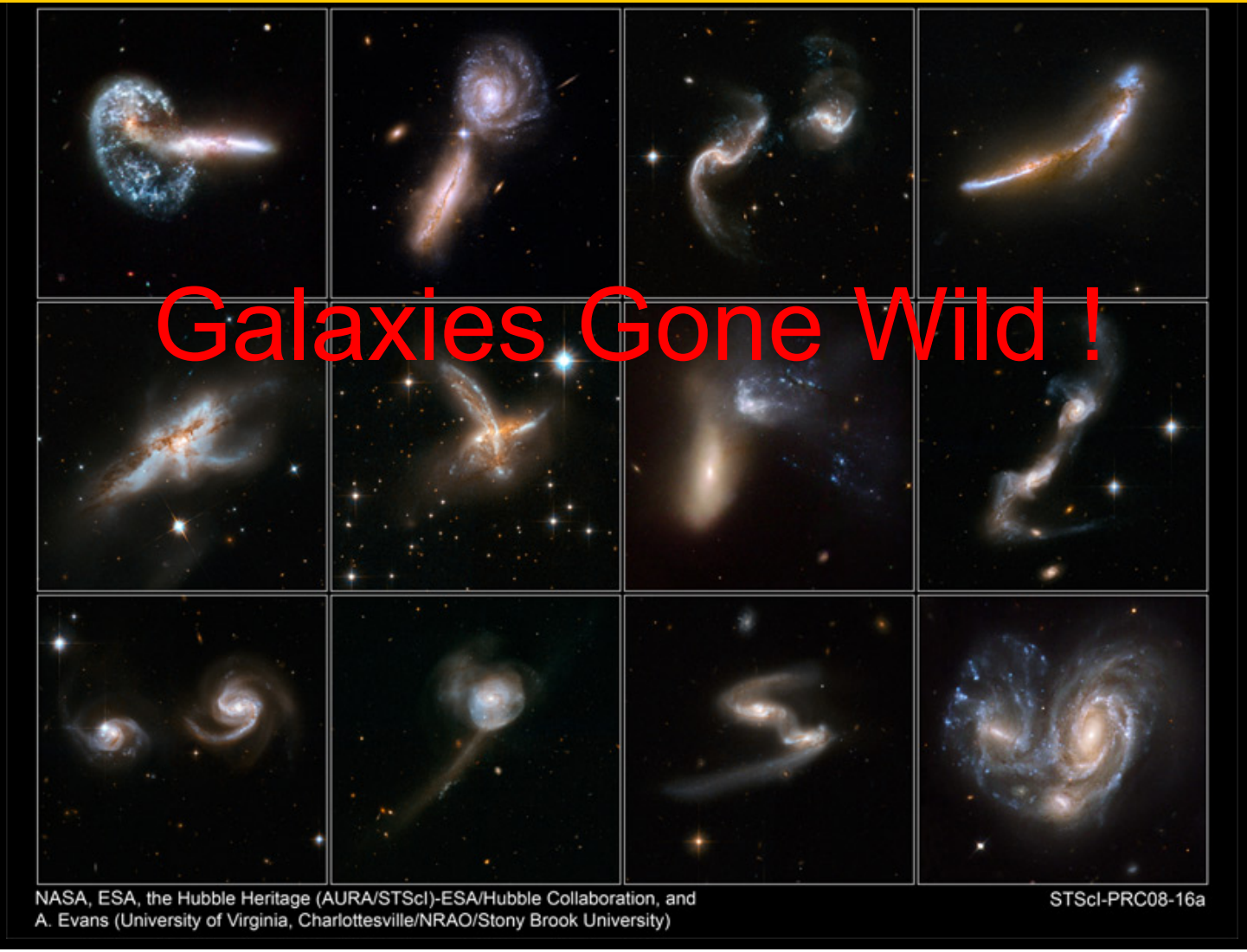
Outline

- The Emergence of Informatics Research in the Sciences
- Human Computation
- Citizen Science: Introduction
- Citizen Science: Examples
- The Zooniverse Project
- **The Zooniverse Machine Learning Challenges**

Key Feature of Zooniverse:
Data mining from the volunteer-contributed labels

- The volunteer's tags, labels, annotations produce a new data flood!
- Apply Machine Learning (data mining) algorithms to learn from those tags
  - How do the volunteer-contributed tags, labels, and annotations correlate with scientist-measured science parameters (in project databases)
- Ultimate goal: Train the automated pipeline classifiers with:
  - Improved classification algorithms
  - Better identification of anomalies
  - Fewer classification errors
  - Based upon millions of training examples …
  - … and hundreds of millions of new examples with class labels
- Additional uses … Statistics deluxe! …
  - User assessment (see paper: http://arxiv.org/abs/0909.2925)
  - Uncertainty Quantification (UQ)
  - Classification certainty vs. Classification dispersion

First Case Study: test SDSS science catalog attributes to find which attributes correlate most strongly with user-classified mergers.



Sloan Science Database Attributes tested

Attribute	Description
$petroMag_{ug}$	Petrosian magnitude colors. A color was calculated for four independent pairs of bands in SDSS (u, g, r, i, z).
$petroRad_u * z$	Petrosian radius, transformed with redshift to be distance-independent.
$invConIndx_u$	Inverse concentration index. The ratio of the 50% Petrosian magnitude to the 90% Petrosian magnitude.
$isoRowcGrad\_u * z$	Gradient of the isophotal row centroid, transformed with redshift to be distance-independent.
$isoColcGrad_u * z$	Gradient of the isophotal column centroid, transformed with redshift to be distance-independent.
$isoA_u * z$	Isophotal major axis, transformed with redshift to be distance-independent.
$isoB_u * z$	Isophotal minor axis, transformed with redshift to be distance-independent.
$isoAGrad_u * z$	Gradient of the isophotal major axis, transformed with redshift to be distance-independent.
$isoBGrad_u * z$	Gradient of the isophotal minor axis, transformed with redshift to be distance-independent.
$isoPhiGrad_u * z$	Gradient of the isophotal orientation, transformed with redshift to be distance-independent.
$texture_u$	Measurement of surface texture.
$lnLExp_u$	Log-likelihood of exponential profile fit.
$lnLDeV_u$	Log-likelihood of De Vaucouleurs profile fit.
$fracDev_u$	Fraction of the brightness profile explained by the De Vaucouleurs profile.

Results of Decision Tree Information Gain analysis

Attribute	Information Gain
$lnLExp_g$	0.10118
$texture_g$	0.07335
$lnLDeV_g$	0.06864
$petroMag_{gr}$	0.06626
$isoAGrad_u * z$	0.05729

Results of cluster separation analysis

Best Separation in Single Dimension	Best Separation Among 1014 Combinations
$isoAGrad_u * z$	$isoAGrad_u * z$
$petroRad_u * z$	$petroRad_u * z$
$texture_u$	$texture_u$
$isoA_z * z$	$isoA_z * z$
$lnLExp_u$	$lnLExp_u$
$lnLExp_g$	$lnLExp_g$
$isoA_u * z$	$petroRad_u * z$, $isoB_z * z$, $isoBGrad_u * z$, $lnLExp_g$
$isoB_z * z$	$isoAGrad_u * z$, $lnLExp_g$
$isoBGrad_u * z$	$petroRad_u * z$, $isoA_u * z$, $isoB_z * z$, $lnLExp_g$
$isoAGrad_z * z$	$isoAGrad_u * z$, $isoBGrad_u * z$, $lnLExp_g$

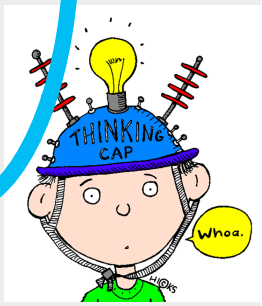
Sloan Science Database Attributes found !!

Attribute	Description
$petroMag_{ug}$	Petrosian magnitude colors. A color was calculated for four independent pairs of bands in SDSS (u, g, r, i, z).
$petroRad_u * z$	Petrosian radius, transformed with redshift to be distance-independent.
$invConIndx_u$	Inverse concentration index. The ratio of the 50% Petrosian magnitude to the 90% Petrosian magnitude.
$isoRowcGrad\_u * z$	Gradient of the isophotal row centroid, transformed with redshift to be distance-independent.
$isoColcGrad_u * z$	Gradient of the isophotal column centroid, transformed with redshift to be distance-independent.
$isoA_u * z$	Isophotal major axis, transformed with redshift to be distance-independent.
$isoB_u * z$	Isophotal minor axis, transformed with redshift to be distance-independent.
$isoAGrad_u * z$	Gradient of the isophotal major axis, transformed with redshift to be distance-independent.
$isoBGrad_u * z$	Gradient of the isophotal minor axis, transformed with redshift to be distance-independent.
$isoPhiGrad_u * z$	Gradient of the isophotal orientation, transformed with redshift to be distance-independent.
$texture_u$	Measurement of surface texture.
$lnLExp_u$	Log-likelihood of exponential profile fit.
$lnLDeV_u$	Log-likelihood of De Vaucouleurs profile fit.
$fracDev_u$	Fraction of the brightness profile explained by the De Vaucouleurs profile.

Results of Decision Tree Information Gain analysis

Attribute	Information Gain
$lnLExp_g$	0.10118
$texture_g$	0.07335
$lnLDeV_g$	0.06864
$petroMag_{gr}$	0.06626
$isoAGrad_u * z$	0.05729

Correlation Zoo !



Results of cluster separation analysis

Best Separation in Single Dimension	Best Separation Among 1014 Combinations
$isoAGrad_u * z$	$isoAGrad_u * z$
$petroRad_u * z$	$petroRad_u * z$
$texture_u$	$texture_u$
$isoA_z * z$	$isoA_z * z$
$lnLExp_u$	$lnLExp_u$
$lnLExp_g$	$lnLExp_g$
$isoA_u * z$	$petroRad_u * z$, $isoB_z * z$, $isoBGrad_u * z$, $lnLExp_g$
$isoB_z * z$	$isoAGrad_u * z$, $lnLExp_g$
$isoBGrad_u * z$	$petroRad_u * z$, $isoA_u * z$, $isoB_z * z$, $lnLExp_g$
$isoAGrad_z * z$	$isoAGrad_u * z$, $isoBGrad_u * z$, $lnLExp_g$

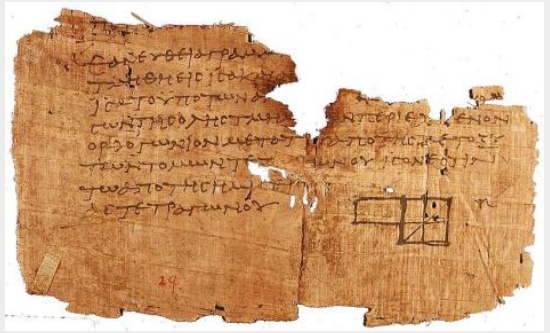
Combinatorial Explosion !!

ML Challenge Problems

Research Awards

Zooniverse Data Mining (Machine Learning) Challenge Problems (2011-2013) – anticipated plans:

- Similar to KDD cups (but no cash award, just fame!)
- Announcement and Challenge issued to ML community
- Seeking algorithmic solutions:
  - How to train and improve science data pipeline processing algorithms using volunteer-contributed tags
  - How to measure "figure of merit" for different users
  - How to apply "figure of merit" to classifications
  - How to characterize and detect anomalies better
  - … [TBD]



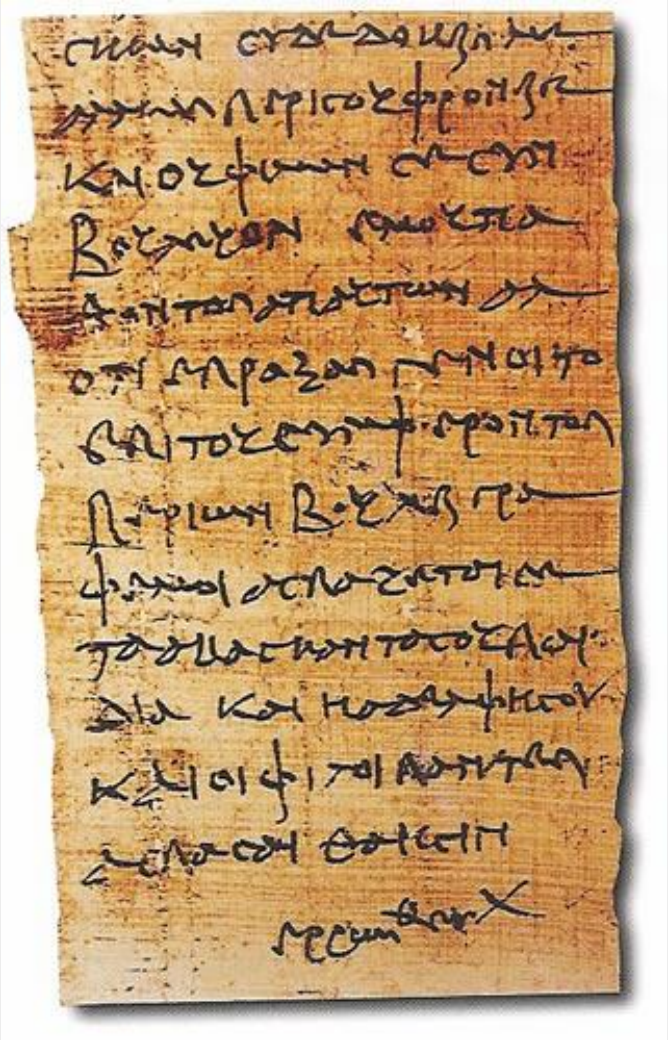
- (Possible) First project (2011): The Digital Papyri Project (probably)
  - Character Recognition: improved automated classification algorithms
  - Sequence Mining: algorithms for mining patterns and sequences in the text (*e.g.,* is this an example of Homer's writings? is this a known writing? what is the context of the writing – property sale, political document, love letter, …?)

ZOONIVERSE
REAL SCIENCE ONLINE

http://zooniverse.org/

Zooniverse Machine Learning Challenge #1:

Stay tuned for announcement later this year !



MS 2650
Bible: Matthew. Egypt, 1st half of 4th c.
Unique text of the Gospel. 8 chapters are the earliest known of this part of the Bible